

Regionale Kontextdaten mit dem Mikrozensus

Pforr, Klaus

Veröffentlichungsversion / Published Version

Arbeitspapier / working paper

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:

GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:

Pforr, K. (2021). *Regionale Kontextdaten mit dem Mikrozensus*. (GESIS Papers, 2021/02). Köln: GESIS - Leibniz-Institut für Sozialwissenschaften. <https://doi.org/10.21241/ssoar.71319>

Nutzungsbedingungen:

Dieser Text wird unter einer CC BY Lizenz (Namensnennung) zur Verfügung gestellt. Nähere Auskünfte zu den CC-Lizenzen finden Sie hier:

<https://creativecommons.org/licenses/by/4.0/deed.de>

Terms of use:

This document is made available under a CC BY Licence (Attribution). For more Information see:

<https://creativecommons.org/licenses/by/4.0>

Regionale Kontextdaten mit dem Mikrozensus

Klaus Pforr

GESIS Papers 2021|02

Regionale Kontextdaten mit dem Mikrozensus

Klaus Pforr

GESIS Papers

GESIS – Leibniz-Institut für Sozialwissenschaften
Dauerbeobachtung der Gesellschaft
German Microdata Lab
Postfach 12 21 55
68072 Mannheim

Telefon: (0621) 12 46 - 0
E-Mail: klaus.pforr@gesis.org

ISSN:	2364-3781 (Online)
Herausgeber, Druck und Vertrieb:	GESIS – Leibniz-Institut für Sozialwissenschaften Unter Sachsenhausen 6-8, 50667 Köln

1 Einleitung

Kleinräumige Kontextdaten sind für viele sozialwissenschaftliche Fragestellungen relevante Einflussfaktoren (Friedrichs & Nonnenmacher, 2014; Galster, 2012; Ham et al., 2012). In der Familiensoziologie wird die Menge potenziell verfügbarer Partner im eigenen Umfeld als Gelegenheitsstruktur für den Partnermarkt und damit für das Entstehen neuer Partnerschaften verstanden (Eckhard & Stauder, 2016, 2018; Stauder, 2011, 2015). In der Bildungs- und Ungleichheitssoziologie spielen die in einer Nachbarschaft vorliegenden Ressourcen wie Kindergärten und Schulen, aber auch Spielgelegenheiten und Verkehrsdichte eine wichtige Rolle für die kindliche Entwicklung (Agostinelli et al., 2020; Bischoff & Tach, 2020; Burdick-Will et al., 2020; Chetty et al., 2014, 2016; Chetty & Hendren, 2018a, 2018b; Hermansen et al., 2020; Weßling, 2016). In der Umweltsoziologie konnte festgestellt werden, dass lokale Umweltbelastungen wie Luftverschmutzung und Verkehrslärm dazu führen, dass einkommensreichere Haushalte eher wegziehen (Best & Rüttenauer, 2018; Currie et al., 2020; Diekmann & Meyer, 2010; Kohlhuber et al., 2006). Hier wurden auch Effekte der Umweltbelastung auf Gesundheit und schulische Leistungen gefunden (Heissel et al., 2019; Pestel & Wozny, 2019). In der Kriminalsoziologie werden verschiedene Nachbarschaftsmerkmale wie lokale Armut als wichtiger Einflussfaktor von Jugenddelinquenz untersucht (Bernasco et al., 2017; Billings et al., 2016; Braga et al., 2017; Legewie, 2018; Legewie & Schaeffer, 2016; Oberwittler et al., 2013; Süß, 2020; Wickes & Hipp, 2018). In der Arbeitssoziologie werden Effekte der lokalen Infrastruktur auf Arbeitsmarktchancen untersucht (Åslund et al., 2017; Dauth & Haller, 2020; Heuermann & Schmieder, 2018; Kawabata & Abe, 2017; Möller & Zierer, 2018), aber auch der Einfluss lokaler Massenentlassungen (Vom Berge & Schmillen, 2015). In der Migrationssoziologie wird ethnische Segregation in Nachbarschaften untersucht, d.h. kleinräumige Kontexte sind hier die abhängige Variable (z.B. Bruch & Mare, 2006; Clark, 1991; Reardon & Firebaugh, 2002). In der politischen Soziologie wird unter dem Schlagwort „Economic Voting“ der Einfluss der lokalen wirtschaftlichen Lage auf die Wahlabsicht untersucht (z.B. González & Granic, 2020; Lewis-Beck & Stegmaier, 2000; Margalit, 2019), und im Nachgang der „Europäischen Flüchtlingskrise“ wird der Einfluss von regionalen Verteilungen von Geflüchteten auf Wahlen betrachtet (z.B. Dustmann et al., 2019; Halla et al., 2017).¹

Voraussetzungen für Verknüpfung von Kontextdaten

Voraussetzung für die Nutzung räumlicher Kontextdaten ist bei den meisten Fragestellungen, dass räumliche Kontextdaten an Daten auf Individualebene zugespielt werden können. Wenn dies nicht möglich ist, ist man entweder auf die Analyse von Zusammenhängen auf der räumlichen Aggregatebene angewiesen, oder man muss das subjektiv erlebte räumliche Nahumfeld der Befragung miterheben. Während letzteres das Problem des ökologischen Fehlschlusses prinzipiell reduzieren kann, ist direkte Messung subjektiv erlebter kleinräumiger Kontexte mit Messfehlern verbunden, und der Zusammenhang zwischen objektiven Kontexten und der subjektiven Wahrnehmung nicht mehr trennbar (Horr, 2016).

Um Kontextdaten an Individualdaten anspielen zu können, müssen in den Individualdaten räumliche Identifikatoren vorliegen, wie z.B. Geo-Koordinaten des Wohnorts oder das 100m-Grid (INSPIRE Thematic Working Group Coordinate Reference Systems & Geographical Grid Systems, 2014), die Gemeinde oder vergleichbaren räumliche Einheiten, in denen der Wohnort liegt. In Deutschland ist aus Datenschutzgründen das Anspielen auf Geokoordinaten-Ebene in der Regel nur unter

¹ Die Auflistung ist eine unvollständige Auswahl von Themenfeldern und Arbeiten, die stellvertretend die Breite und Aktualität der Bedeutung von Kontextdaten belegen soll.

eingeschränkten Bedingungen möglich (Bensmann et al., 2020; Bluemke et al., 2017; Kroll & Schnell, 2016; Schweers et al., 2016). Darüber hinaus benötigt man messfehlerarme Kontextdaten auf möglichst kleinräumiger Ebene. Die möglichst kleinräumige Messung von Kontextdaten ist aus mehreren Gründen notwendig. Erstens kann bestenfalls auf der kleinsten räumlichen Ebene verknüpft werden, auf der die Kontextdaten vorliegen und auf der bei den Individualdaten räumliche Identifikatoren verfügbar sind. Wenn z.B. bei den Individualdaten auf Gemeindeebene verknüpft werden kann, die Kontextdaten aber auf Kreisebene vorliegen, müssen die Gemeinden auf Individualdatenebene mit dem damit einhergehenden Informationsverlust aggregiert werden. Zweitens muss abhängig von der Fragestellung das räumliche Messniveau dem unterstellten Mechanismus entsprechen (Modifiable Area Unit Problem (MAUP): Lee et al., 2008; Openshaw, 1983). Wenn man den Zusammenhang zwischen dem Migrant*innenanteil in Nachbarschaften und der Wohnortwahl beeinflusst, Menschen bei der Bewertung aber nur die umliegenden Straßenzüge berücksichtigen, weiter entfernte Häuser im selben amtlichen definierten Stadtteil (Behrens et al., 2019, S. 162 ff.) dabei aber außer Acht lassen, sind Kontextdaten auf Stadtteilebene und vermutlich auch auf 100m-Grid-Ebene zu ungenau, um die kleinräumigen Effekte zu messen. Durch die Aggregation sind solche Kontextdaten mit so großem Messfehler behaftet, dass Zusammenhänge stark unterschätzt werden.

Verknüpfungsmöglichkeiten in Deutschland²

In Deutschland ist es bei den meisten großangelegten Umfrageprogrammen seit langem möglich, Kontextdaten auf Gemeindeebene zuzuspielen. In den letzten Jahren wurde die Zuspielmöglichkeit auf 100m-Grid-Ebene und vereinzelt auf Geo-Koordinaten-Ebene ausgeweitet. Demgegenüber steht das Angebot an Kontextdaten, das an Umfragedaten anspielt werden kann. Aus amtlichen Quellen erhält man für Gesamtdeutschland aus der INKAR-Datenbank und der Regionaldatenbank Daten auf Kreis- und Gemeindeebene. Aus dem Zensus 2011 erhält man Daten auf 100m-Grid-Ebene. Aus den Datenbanken des Verbands deutscher Städtestatistiker über Daten der Kommunalstatistik erhält man Daten auf Gemeindeebene. Neben diesen aggregierten amtlichen Datenbeständen gibt es mit dem Mikrozensus und den Registerdaten der deutschen Rentenversicherung und des IAB amtliche Mikrodaten, die grundsätzlich als regionale Kontextdaten genutzt werden können. Die Daten des Mikrozensus liefern Daten prinzipiell auf Gemeindeebene, die Daten der Rentenversicherung und des Instituts für Arbeitsmarkt- und Berufsforschung (IAB) der Bundesagentur für Arbeit (BA) auf Kreisebene, wobei, seit 2018 auch im Mikrozensus die Geokoordinaten der befragten Haushalte erfasst werden, für alle drei Quellen grundsätzliche eine beliebige Genauigkeit möglich ist. Zu den inhaltlichen Daten aus amtlichen Quellen kommen die topographischen Daten aus amtlichen Quellen hinzu. Aus den Daten des Bundesinstituts für Bau-, Stadt- und Raumforschung (BBSR) und des Bundesamts für Kartographie und Geodäsie (BKG) sind Daten über alle amtlichen definierten Regionaleinheiten verfügbar und über den Baubestand in Deutschland. Diesen amtlichen Daten stehen auf verschiedene Art erfasste, im weitesten Sinne physikalische Daten gegenüber. Diese beinhalten z.B. Wetter-, Sonneneinstrahlungs-, Verkehrslärm-, Bodenbedeckungs- und Lichtabstrahlungsdaten. Quellen sind hier z.B. der Deutsche Wetterdienst oder das Bundesumweltamt. Diese Daten sind grundsätzlich in der Regel mit Geokoordinaten versehen, d.h. räumlich auf beliebig kleiner Ebene verfügbar, wobei zu beachten ist, dass die Wetterdaten in der Rohform an bestimmten Messstationen erfasst werden und Verkehrsdaten ausgehend von der Straßenverläufen, beobachteten Verkehrsströmen und Modellen aus der physikalischen Akustik geschätzt werden. Daneben gibt es private Anbieter, die räumliche Kontextdaten professionell erstellen und vertreiben. Z.B. bietet BIK Aschpurwis + Behrens GmbH unter anderem regionale

² Eine Übersicht über verfügbare regionale Datenquellen findet man in den unregelmäßig aktualisierten Regionalen Standards (Behrens et al., 2019).

Daten über Ballungsräume, die aus Pendelverhalten abgeleitet werden, an, und infas 360 GmbH und microm GmbH bieten auf sehr kleinräumiger Ebene, wie z.B. auf Straßenabschnittsebene, Daten über soziale und sozio-ökonomische Merkmale an, die aus verschiedenen Quellen erstellt werden. Den Daten der Privatanbieter ist gemein, dass die Quellen und die Erstellung nicht nachvollziehbar sind. Ferner kann man aus dem Datenbestand der OpenStreetMap (OSM) Daten über die räumliche Verteilung von Dienstleistungseinrichtungen, wie z.B. Kindergärten, Spielplätze oder Restaurants, erstellen. Da die OSM im crowdworking-Verfahren gepflegt wird, ist die Datenqualität schwer einschätzbar. Schließlich können räumliche Kontextdaten grundsätzlich auch aus konventionellen Befragungen wie dem ALLBUS geschätzt werden. Durch die geringen Fallzahlen sind hier Schätzungen mit tolerablen Varianzen bestenfalls auf Bundeslandebene möglich.

Potential des Mikrozensus

Wie oben festgestellt, sind möglichst messfehlerarme Kontextdaten auf möglichst kleiner Ebene eine Voraussetzung für das Zusammenspielen. Dass Kontextdaten möglichst messfehlerarm sein sollten, erscheint trivial, ist aber angesichts der oben dargestellten verfügbaren Kontextdaten in der Praxis relevant. Die Übersicht der verfügbaren Kontextdaten ergibt, dass in aggregierten amtlichen Daten Kontextdaten auf Gemeindeebene grundsätzlich verfügbar sind, die Indikatoren häufig aber schlechte Proxyvariablen für die eigentlich interessierenden Variablen sind. Bei physikalischen Daten hat man eine räumliche feinere Messung, aber hier sind verfügbare Indikatoren bei vielen Fragestellungen noch weiter von den interessierenden Konstrukten entfernt. Die Daten der Privatanbieter bieten zwar teilweise interessante Indikatoren auf kleinräumiger Ebene, aber die Erstellung und die Quelle der Ausgangsdaten sind vollkommen unklar. Demgegenüber stehen die sozialwissenschaftlichen Umfragen, in denen man bei den meisten Fragestellungen zufriedenstellende Messungen findet, aber die räumliche Ebene Bundesland nicht akzeptabel ist. Ein Kompromiss zwischen beiden Problemquellen stellen die amtlichen Mikrodaten, insbesondere der Mikrozensus, dar. Mit dem Mikrozensus kann grundsätzlich eine breite Menge von Indikatoren abgeleitet werden, die prinzipiell auf Gemeindeebene und mit der Einführung der Geokoordinaten der Haushalte potenziell auf kleinerer Ebene verfügbar ist.

Vor dem Hintergrund, dass gute räumliche Kontextdaten auf kleinräumiger Ebene höchst relevant sind, und der Mikrozensus hierfür ein guter Ausgangspunkt darstellt, soll in diesem Papier gezeigt werden, wie aus dem Mikrozensus räumliche kontinuierliche regionale Kontextdaten geschätzt werden können, die anderen Datenquellen auf beliebig kleiner Ebene zugespielt werden können. Im zweiten Teil wird eine prototypische Schätzung mit dem MZ 2000 Regionalfiler gezeigt. Im dritten Kapitel werden die Erweiterungsarbeiten mit dem MZ 2015 im On-Site-Zugang dokumentiert. Das vierte Kapitel zeigt eine resümierende Analyse über das Potential von Kontextdaten auf unterschiedlichen Raumeinheiten.

2 Schätzung kontinuierlicher räumlicher Kontextdaten mit dem Mikrozensus Regionalfile 2000

Zunächst soll die Ausgangslage für die Schätzung einer kontinuierlichen räumlichen Verteilung aus dem Mikrozensus erläutert werden. Im Mikrozensus liegen die Daten in regionalen Einheiten mit amtlich definierten Grenzen zusammengefasst vor. Die kleinste räumliche Einheit sind Gemeinden. Ziel ist die Schätzung der darunterliegenden, realen, kontinuierlichen, räumlichen Verteilung der im Mikrozensus enthaltenen Merkmale. Mit der geschätzten kontinuierlichen räumlichen Verteilung können räumliche Kontextdaten prinzipiell auf beliebig kleiner Ebene erzeugt werden.

Zur Schätzung der räumlichen Verteilung wird ein besonderes Kerndichteschätzer-Verfahren von Groß et al. (2017) verwendet. Das Verfahren von Groß et al. betrachtet das Schätzproblem als Messfehlerproblem, das entsteht, wenn eine kontinuierliche Variable in Gruppen zusammengefasst wird. In einer ersten Anwendung schätzen Groß und Rendtel (2016) aus einer klassierten Einkommensvariable die zugrundeliegende kontinuierliche Einkommensverteilung. Für die Schätzung der räumlichen Verteilung wird das Verfahren vom eindimensionalen Fall auf die zwei Raumdimensionen erweitert. Die besondere Eigenschaft des Verfahrens von Groß et al. ist, dass die Verteilung die bekannten Anteile in den aggregierten Gruppen wiedergibt.

Aus praktischen Gründen wird das Verfahren zunächst mit den Daten des Mikrozensus-Regionalfile 2000 (Christians & Wirth, 2009; Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder, 2018) getestet, da diese Daten als Scientific-Use-File genutzt werden können. Im Mikrozensus-Regionalfile ist die kleinste räumliche Ausgangseinheit die „Mikrozensuskreisregion“. Diese künstlich gebildeten Raumeinheiten sind eine Menge von Kreisen und Zusammenfassungen von Kreisen, die so gebildet wurden, dass die Fallzahlen im MZ Regionalfile innerhalb der Mikrozensuskreisregionen groß genug sind, um dem Kriterium der faktischen Anonymität zu genügen. Der Unterschied zwischen den Kreisen und den Mikrozensuskreisregionen ist in Abbildung 1 dargestellt. Leider gibt es das Mikrozensus-Regionalfile nur für das Erhebungsjahr 2000.

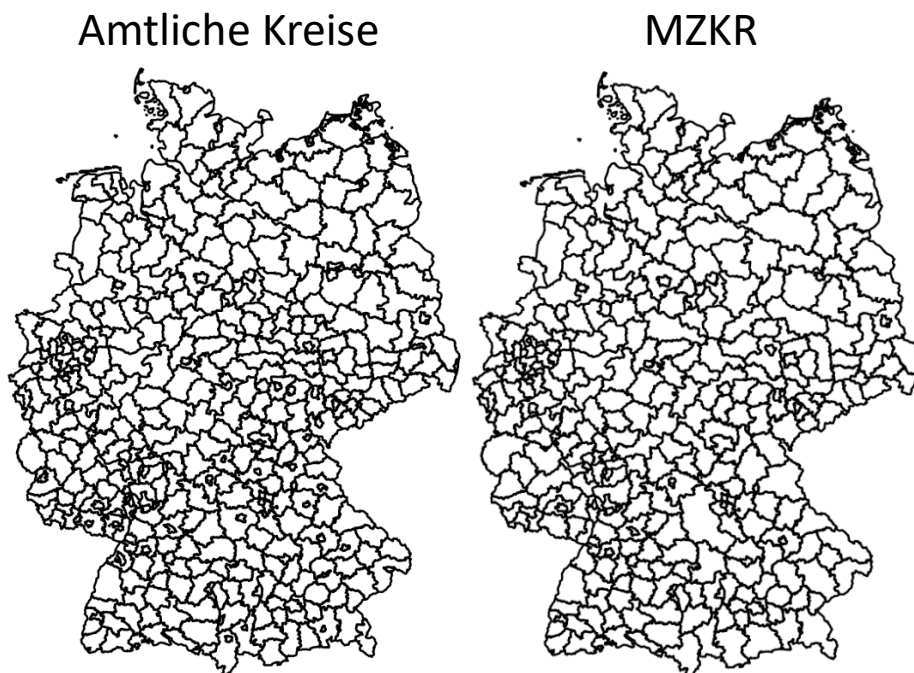


Abbildung 1 Unterschied zwischen Mikrozensuskreisregionen und Kreisen

2.1 Aufbereitung und Schätzung

Aus dem Mikrozensus-Regionalfile werden zunächst folgende Variablen ausgewählt, aus denen räumliche Verteilungen geschätzt werden sollen:

Tabelle 1 Verwendete Variablen aus dem Mikrozensus-Regionalfile

Variablenname	Variablenlabel
ef1	Land der Bundesrepublik
ef3s	Systemfreie Auswahlbezirksnummer
ef4s	Systemfreie Lfd. Nr. des Haushalts im Auswahlbezirk
ef5s	Systemfrei Lfd. Nr. der Person im Haushalt - aus der Erhebungsliste
ef30	Alter
ef32	Geschlecht
ef39	Lebenspartner der 1. Person
ef43	Deutsche Staatsangehörigkeit
ef44	1. Staatsangehörigkeit
ef72	Art der besuchten Schule
ef83	Zeitaufwand für den Hinweg zur Schule
ef286	Allgemeiner Schulabschluss vorhanden
ef287	Höchster allgemeiner Schulabschluss
ef506	Bevölkerung in Privathaushalten

Aus diesen Variablen werden folgende Variablen konstruiert:³

Tabelle 2 Konstruierte Variablen

Variablenname	Variablenlabel	Ausprägungen	Value label
national	Erste Staatsbürgerschaft	1	Deutsch
		2	Türkisch
		3	Italienisch
		4	Ex-Jugoslawisch
		5	GUS
		6	Griechisch
		7	Polnisch
		8	Sonstige
schulabschl	Höchster Schulabschluss (nur Schüler)	1	Kein/Hauptschule
		2	Realschule/POS
		3	Abitur/Fachabitur/EOS
oberstufe	Klasse 11-13 im Alter von 18	1	Ja
		0	Nein
mobility	Bildungsmobilität	1	18-jähriges Kind ist nicht in Sek-II, mind. ein Elternteil hat Abitur
		2	18-jähriges Kind ist nicht in Sek-II, beide Eltern haben kein Abitur oder 18-jähriges

³ Bei allen Indikatoren beschränke ich mich auf die Bevölkerung in Privathaushalten. Die Aufbereitung findet sich in Code 1 im Anhang.

			Kind ist in Sek-II und mind. ein Elternteil hat Abitur
		3	18-jähriges Kind ist in Sek-II, Eltern haben kein Abitur
homo	Gleichgeschlechtlich kohabierend	1	Gleichgeschlechtlich
		0	gegengeschlechtlich
schulweg	Zeitaufwand für den Hinweg zur Schule (nur Schüler)	1	Unter 10 Minuten
		2	10 bis unter 30 Minuten
		3	1/2 bis unter 1 Stunde
		4	1 Stunde und mehr
hheinkat	Haushaltsnettoeinkommen im April (ohne Selbstständige), Quartile	1	Bis unter 2200 DM
		2	2200 DM bis unter 3500 DM
		3	3500 DM bis unter 5000 DM
		4	5000 DM und mehr

Aus diesen konstruierten Variablen wird dann ein Datensatz erzeugt, der für jede Variable die Anzahl der für die jeweiligen relevanten Einheiten (bei Personenmerkmalen die Anzahl aller Personen, bei Haushaltsmerkmalen die Anzahl aller Haushalte) insgesamt und in einzelnen Ausprägungen der kategorialen Variablen über die Mikrozensuskreisregionen enthält. Zur Veranschaulichung ist in Tabelle 3 ein Ausschnitt des Datensatzes dargestellt.

Tabelle 3 Ausschnitt der aufbereiteten Aggregatdaten

SHN ^a	national_N1 ^b	national_NE1 ^c	national_N2 ^d	national_NE2 ^e	schulabschl_N1 ^f	schulabschl_NE1 ^g
100200000 1708	1830	64	1830	611	1393	
100300000 1579	1717	41	1717	636	1272	
105100000 1031	1065	0	1065	453	758	
105300000 1303	1359	26	1359	476	925	
105400000 1298	1338	3	1338	532	954	
105500000 1632	1658	0	1658	607	1217	
105600000 2276	2403	45	2403	777	1763	
105700000 984	999	0	999	371	691	
105800000 2566	2661	33	2661	912	1930	
105900000 1990	2092	10	2092	782	1479	
106000000 1759	1858	25	1858	639	1313	
106100000 1067	1113	14	1113	443	770	
106200000 1629	1699	4	1699	491	1185	
200000000 11597	13355	466	13355	3968	9525	

Anmerkung: a) 11-stellige statistische Schlüsselzahl, b) Anzahl Personen mit 1. Staatsbürgerschaft deutsch, c) Anzahl Personen, d) Anzahl Personen mit 1. Staatsbürgerschaft türkisch, e) Anzahl Personen, f) Anzahl Personen, die keine Schüler sind, die keinen oder einen Hauptschulabschluss haben, g) Anzahl Personen, die keine Schüler sind.

In der ersten Spalte steht mit SHN eine 11-stellige eindeutige Kreis-Identifikationsnummer. In der zweiten Spalte steht die Anzahl der Personen mit Deutsch als erster Staatsbürgerschaft in den einzelnen Kreisregionen. In der dritten Spalte steht die Anzahl aller Personen in den Kreisregionen. In den folgenden Spalten stehen die entsprechenden Anzahlen für die anderen Variablen.

Im nächsten Schritt wird aus den aufbereiteten Aggregatdaten und den Shapefiles der Kreise die räumliche kontinuierliche Verteilung geschätzt.⁴ Die Shapefiles der Kreise aus dem Jahr 2000 sind beim Bundesamt für Kartographie und Geodäsie frei verfügbar (GeoBasis-DE & Bundesamt für Kartographie und Geodäsie, 2000). Vor der eigentlichen Schätzung müssen die Shapefiles der Kreise aufbereitet werden. Zunächst werden die Kreisfusionen, wie sie in der Dokumentation „Zusammenfassung der Kreise in Mikrozensuskreisregionen (MZKR)“ von Christians und Wirth (2009) aufgeführt sind, nachvollzogen. Danach werden Shapes, die keine bewohnte Fläche enthalten, gelöscht. Zuletzt werden für die in Tabelle 2 dargestellten Variablen die Anteile über die Fläche Deutschlands geschätzt.

2.2 Geschätzte räumliche Verteilungen

Im Folgenden werden die geschätzten Verteilungen der aufbereiteten Variablen über Deutschland dargestellt. Die Verteilungen der selteneren Nationalitäten GUS, Griechisch, Polnisch und Sonstige werden nicht gezeigt, da die geringen Fallzahlen zu extremen Verteilungsschätzungen führen.

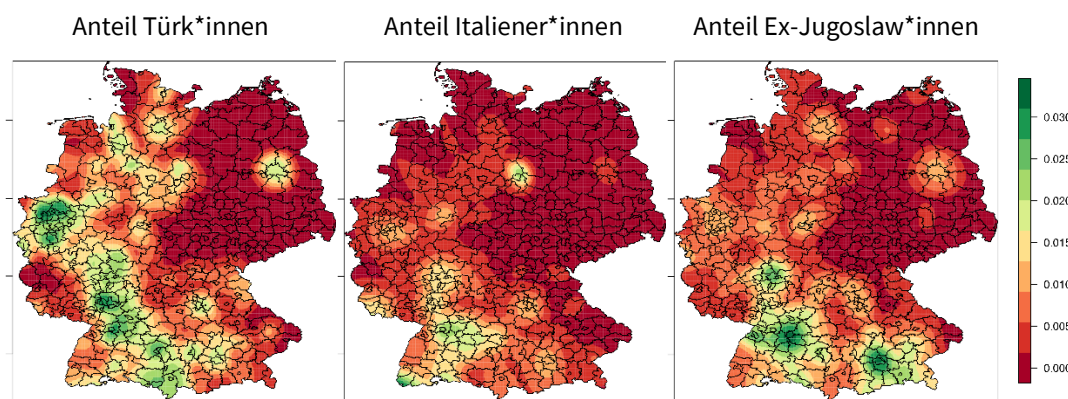


Abbildung 2 Räumliche Verteilung der Anteile von Personen mit türkischer, italienischer und ex-jugoslawischer Staatsbürgerschaft

Abbildung 2 zeigt die räumliche Verteilung der Anteile von Personen mit verschiedenen nichtdeutschen Staatsbürgerschaften im Verhältnis zur Anzahl aller Personen. Aus inhaltlicher Sicht ist erkennbar, dass alle Ausländer*innen-Populationen in 2000 – abgesehen von Berlin – überwiegend in Westdeutschland leben. In Westdeutschland erkennt man Häufungen in großen Industriestädten. Darüber hinaus sind Unterschiede in den Verteilungen zwischen den Ausländer*innengruppen erkennbar. Der Türk*innen-Anteil ist im Ruhrgebiet höher als der der anderen Gruppen, der Anteil der Ex-Jugoslaw*innen ist in München und im Raum um Stuttgart höher, und der Anteil der Italiener*innen ist im Umfeld von Wolfsburg etwas höher als die Anteile der anderen Gruppen.

⁴ Das R-Script findet sich in Code 2 im Anhang.

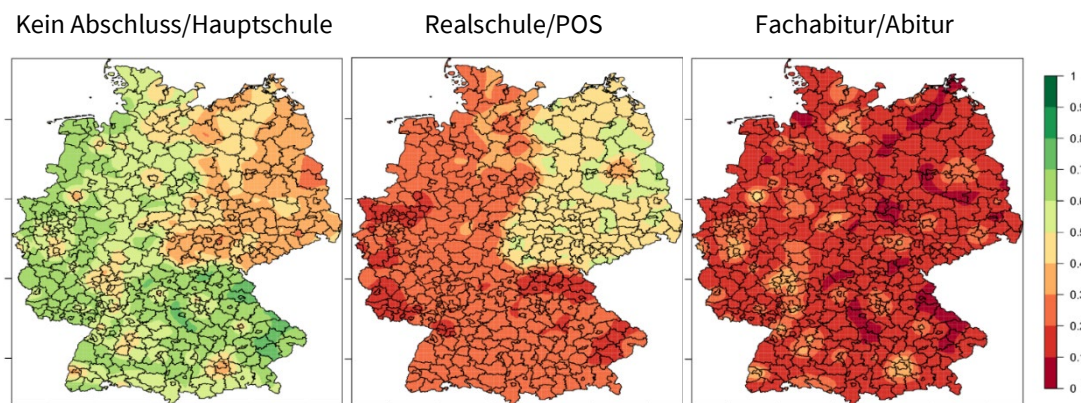


Abbildung 3 Räumliche Verteilung des höchsten erreichten Schulabschlusses

Die räumliche Verteilung des höchsten erreichten Schulabschlusses in Abbildung 3 zeigt auf den ersten Blick Unterschiede zwischen Ost- und Westdeutschland. Während in Ostdeutschland der Anteil der Personen mit höchsten Abschluss Realschule/POS um etwa 20 Prozentpunkte höher ist als im Westen, ist der Anteil von Personen ohne Abschluss oder Hauptschulabschluss um etwa 20 Prozentpunkte höher als im Osten. Bei Fachabitur/Abitur-Anteil ist kein deutlicher Unterschied zwischen Ost- und Westdeutschland erkennbar. Auf regionaler Ebene zeigt sich, dass der Anteil der Personen ohne Abschluss oder mit Hauptabschluss außerhalb der Großstädte höher ist. Dagegen ist in den Großstädten der Anteil der Personen mit Fachabitur oder Abitur höher als in ländlicheren Gebieten.

Eine andere Perspektive auf die räumliche Verteilung von Bildung erlaubt die besuchte Klasse von jungen Schüler*innen. Abbildung 4 zeigt die Verteilung des Anteils der 18-Jährigen, die die Sekundarstufe II besuchen, an allen 18-Jährigen. Es zeigt sich allgemein, dass 18-Jährige in Bayern seltener die Sekundarstufe II besuchen, dagegen 18-Jährige in großen Teilen Nordrhein-Westfalens häufiger.

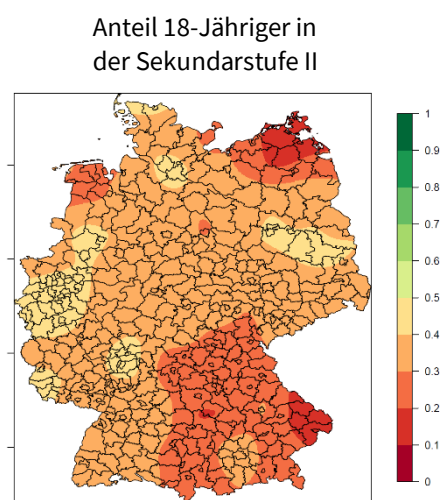


Abbildung 4 Räumliche Verteilung der 18-Jährigen in Sekundarstufe II

Bei zusammenlebenden Familien erlaubt der höchste Schulabschluss der Eltern und die besuchte Schularart der Kinder im Schulalter eine Analyse des Zusammenhangs zwischen elterlicher Bildung und der Bildung des Kindes. Abbildung 5 zeigt die räumliche Verteilung von drei Anteilen, die die Bildungsmobilität widerspiegeln. Der Anteil der Bildungsaufsteiger ist der Anteil an 18-Jährigen in der Sekundarstufe II von Eltern ohne Abitur an allen 18-Jährigen mit Eltern ohne Abitur. Reproduktion bezeichnet den Anteil an 18-Jährigen in der Sekundarstufe II von Eltern mit Abitur⁵ und an 18-Jährigen, die nicht die Sekundarstufe II besuchen, von Eltern ohne Abitur. Bildungsabstieg stellt den Anteil der 18-Jährigen dar, die nicht in der Sekundarstufe II sind und deren Eltern Abitur haben. Die Verteilungen zeigen erstens, dass unabhängig von der räumlichen Verteilung der Reproduktionsanteil sehr viel höher als ist als die Aufstiegs- und Abstiegsanteile. Die regionalen Unterschiede sind gering. Bei der Verteilung von Reproduktion ist ein schwacher Bundeslandunterschied erkennbar. In Nordrheinwestfalen und Niedersachsen ist die Reproduktion schwächer, in Bayern stärker.

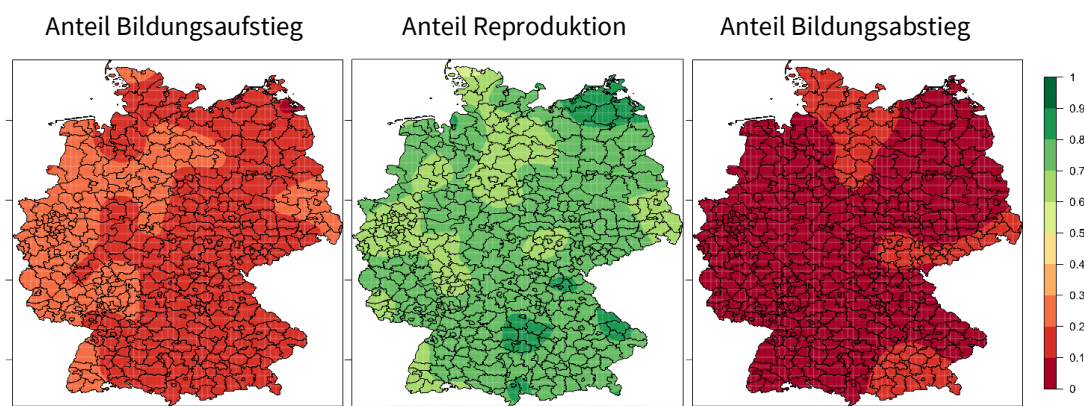


Abbildung 5 Räumliche Verteilung der Bildungsmobilität

Der Haushaltszusammenhang im Mikrozensus erlaubt die Untersuchung von seltenen Formen des Zusammenlebens. Abbildung 6 zeigt die räumliche Verteilung der Anteile der gleichgeschlechtlichen kohabitierenden Paare an allen kohabitierenden Paaren.

⁵ Mit Eltern ohne Abitur ist hier immer gemeint, dass beide Eltern kein Abitur haben, mit Eltern mit Abitur ist gemeint, dass mindestens ein Elternteil Abitur hat.

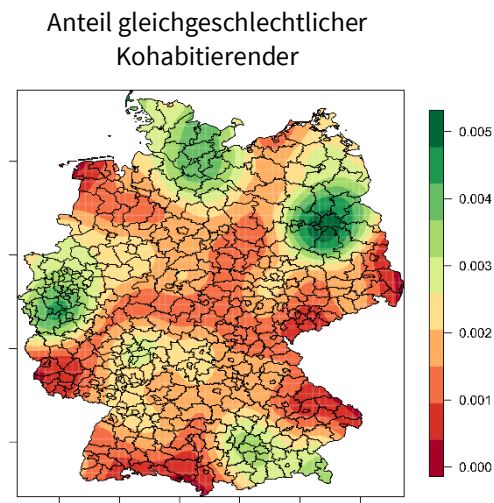


Abbildung 6 Räumliche Verteilung der gleichgeschlechtlichen Kohabitierenden

Zunächst ist zu beachten, dass der Anteil gleichgeschlechtlicher Kohabitierender allgemein klein ist, so dass die Skala im Vergleich zu den anderen Abbildungen geändert wurde. Es zeigt sich eine Konzentration der gleichgeschlechtlichen Kohabitierenden in den Großstädten Berlin, Köln, Hamburg, München und Frankfurt einerseits, und andererseits einige räumliche zusammenhängende Regionen wie z.B. im Saarland, in denen es fast keine gleichgeschlechtlichen Kohabitierenden gibt.

Abbildung 7 zeigt die räumliche Verteilung der Schulweglängen. Man erkennt erstens, dass die Modalschulweglängen in den meisten Regionen zwischen 10 und 30 Minuten liegen, und es überall in Deutschland kaum Schulweglängen von über einer Stunde gibt. Hinsichtlich der regionalen Unterschiede zeigt sich ein unklares Bild. In Teilen Niedersachsens gibt es im Unterschied zu Restdeutschland Gebiete, in denen Schulwege unter 10 Minuten häufiger sind als Schulwege zwischen 10 und 30 Minuten. Für Berlin ist eine generelle relative Erhöhung der Schulweglänge erkennbar.

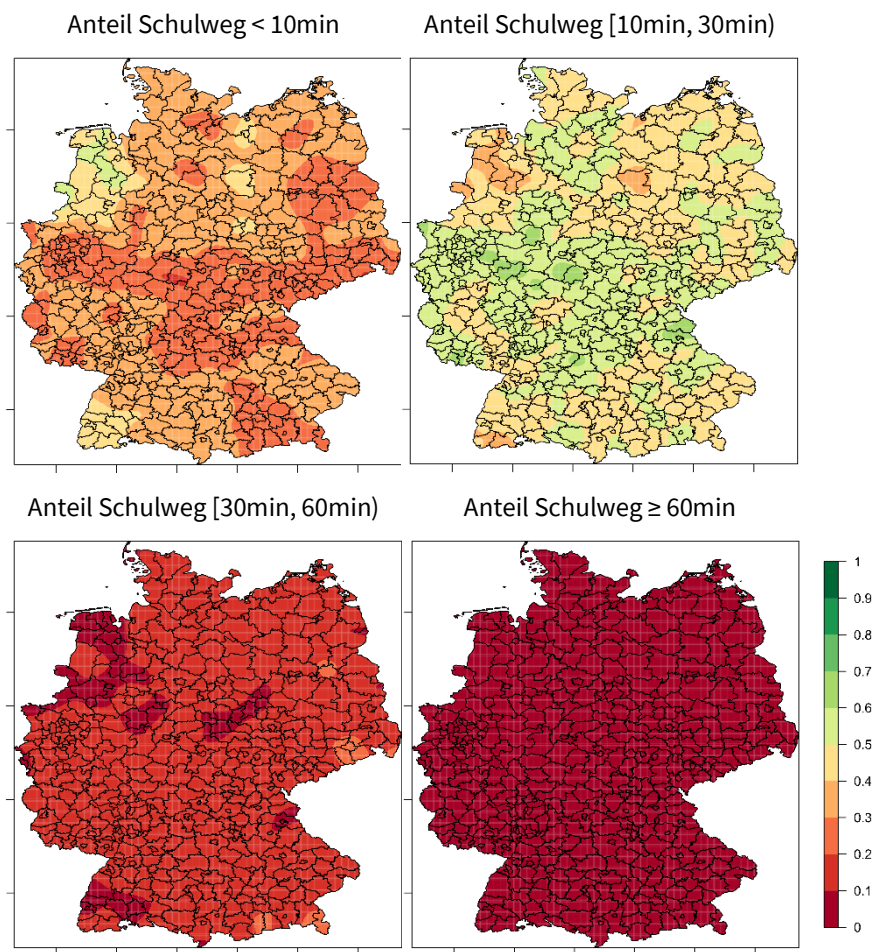


Abbildung 7 Räumliche Verteilung der Schulweglängen

Schließlich kann man mit dem Mikrozensus die räumliche Verteilung des Haushaltseinkommens untersuchen. Abbildung 8 zeigt die Verteilung der Anteile der Haushalte, die in den einzelnen Haushaltsnettoeinkommensquartilen⁶ liegen.

⁶ Da im Mikrozensus das Einkommen in kategorisierter Form vorliegt, werden die Quartile so gebildet, dass die Einkommensgruppe in etwa vier gleich große Gruppen zusammengefasst werden.

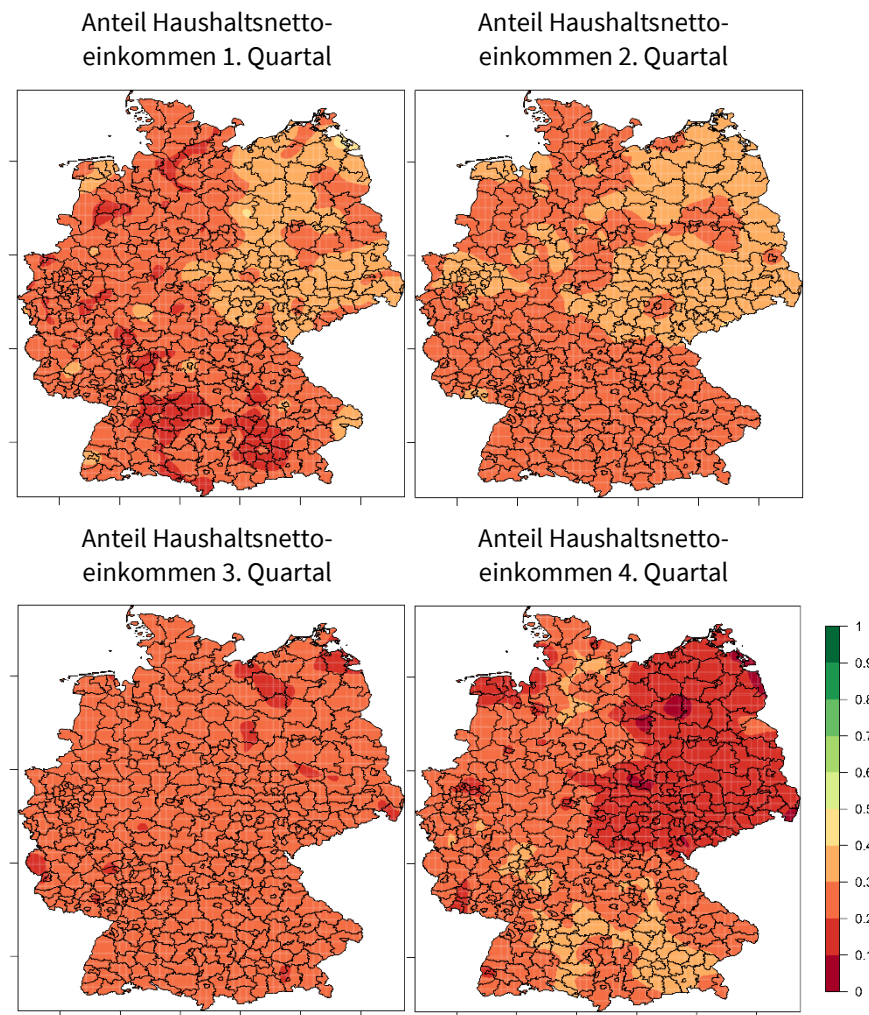


Abbildung 8 Räumliche Verteilung der Verteilung des Haushaltsnettoeinkommens

Die räumlichen Verteilungen zeigen erstens einen Unterschied zwischen Ost- und Westdeutschland. In Ostdeutschland sind abgesehen vom Berliner Umland die Anteile in den unteren beiden Quartilen höher als in Westdeutschland, dagegen sind die Anteile im oberen Quartil niedriger als in Westdeutschland. Darüber hinaus sieht man bei den Großstädten Berlin, München, Stuttgart, Hamburg und Frankfurt Unterschiede zwischen Stadtkreis und den umliegenden Kreisregionen. Hier sind jeweils in den Städten die Anteile der niedrigeren Quartile höher, dagegen im Umland die Anteile der höheren Quartile höher als in der Stadt, aber auch im Vergleich zu Restdeutschland.

2.3 Stichprobenfehler

Das verwendete Verfahren zur Schätzung der räumlichen Verteilung geht eigentlich davon aus, dass die zugrundeliegenden Daten Vollerhebungen der betrachteten Grundgesamtheiten sind. Der Mikrozensus hat zwar eine vergleichsweise hohe Auswahlwahrscheinlichkeit, ist aber nichtsdestot-

rotz nur eine Stichprobe, d.h. die Schätzungen sind von Stichprobenvarianz betroffen. Im Folgenden wird eine einfache Schätzung der Stichprobenvarianz unternommen.⁷

Die Varianz wird mit Hilfe eines Bootstrap-Ansatzes geschätzt. Um den Rechenaufwand in Grenzen zu halten, beschränkt sich die Schätzung auf den Männeranteil in Nordrhein-Westfalen. Da das Geschlecht einigermaßen gleichverteilt ist, und an sich eine geringe Varianz des Männeranteils über die Kreise vorliegt, sollte dieses Merkmal für die Schätzung der Stichprobenvarianz besonders geeignet sein. Zunächst werden aus den Rohdaten des Mikrozensus-Regionalfile 2000 die Männeranteile in den Mikrozensuskreisregionen geschätzt. Im zweiten Schritt werden 50 Bootstrap Samples aus der Menge der Mikrozensuskreisregionen mit Zurücklegen gezogen. Mehrfach gezogene Kreisregionen werden danach aggregiert, sodass in den einzelnen Bootstrap Samples jede Kreisregion nur einmal vorkommt. Dann werden in R zuerst die Shapefiles vorverarbeitet, wobei hier der Schritt des Zusammenfassens einzelner Kreise aus Anonymisierungsgründen wegfällt, da in Nordrhein-Westfalen keine Kreise zusammengefasst werden mussten. Danach werden aus den Koordinaten der Zentroide der Kreisregionen und der Anzahl der Männer und der Gesamtpersonenzahl in den einzelnen Kreisregionen der jeweiligen Bootstrap Samples 50 Datensätze erzeugt. Aus diesen 50 Datensätzen werden mit dem Kerndichteschätzer 50 Anteile in den Kreisregionen und 50 räumliche Verteilungen des Männeranteils über die Fläche Nordrhein-Westfalens geschätzt. Aus den 50 Männeranteilen über die 50 Bootstrap Samples wird die Varianz als die Differenz zwischen dem 95%- und 5%-Perzentil der realisierten Verteilung der Anteile über die Kreisregionen geschätzt und auf die Kreisregionen geplottet. Schließlich wird aus den 50 räumlichen Verteilungen der Männeranteile über die Fläche die Varianz analog wie vorher über die Fläche geschätzt und geplottet. In Abbildung 9 sind die so geschätzten Varianzen des Männeranteils dargestellt.

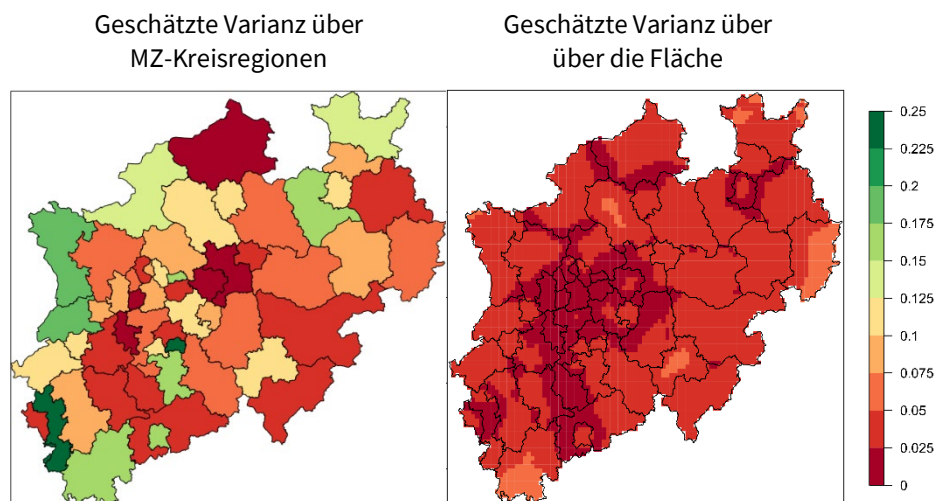


Abbildung 9 Geschätzte Stichprobenvarianz der räumlichen Verteilung des Männeranteils in Nordrhein-Westfalen

Die Abbildung zeigt, dass die geschätzte Stichprobenvarianz über die Fläche sehr viel geringer ist als die Varianz der Anteile über die Mikrozensus-Kreisregionen. Die Reduktion folgt daraus, dass der Schätzalgorithmus die räumliche Kovarianz des Männeranteils über die Kreisregionen ausnutzt. Dies ist daran erkennbar, dass die Stichprobenvarianzen über die Fläche in den Grenzgebiete-

⁷ Der Code zur Aufbereitung in Stata findet sich in Code 3 im Anhang. Der Code für die Schätzung und Visualisierung in R findet sich in Code 4 im Anhang.

ten der einzelnen Kreisregionen besonders gering sind, da hier die Anteile aus mehreren Kreisregionen verwendet werden.

2.4 Zwischenfazit

Aus den geschätzten räumlichen Verteilungen und der geschätzten Varianz der Verteilungen lässt sich zusammenfassend feststellen, dass das Verfahren von Groß et al. (2017) mit dem Mikrozensus-Regionalfile 2000 großflächige räumliche Strukturen abbilden kann. Die Sampling Varianzen der Punktschätzer, die sich daraus ergeben, dass der Mikrozensus eine Stichprobe ist, erscheinen zufriedenstellend, wobei das hier präsentierte Schätzverfahren nur eine prototypische Abschätzung darstellt. Die räumlichen Verteilungen der Varianzschätzungen produzieren durch den Algorithmus kleinere Varianzen an den Binnengrenzen, dagegen höhere Varianzen an den Außengrenzen und im Inneren von großen räumlichen Einheiten. Davon ausgehend ist zu erwarten, dass die geschätzten Verteilungen stark von den gegebenen räumlichen Einheiten abhängen, d.h. auch hier sollte das Modifiable Area Unit Problem (Openshaw, 1983) zum Tragen kommen. Im Abschnitt 4 soll geschätzt werden, inwieweit die Schätzungen von den hier vorliegenden räumlichen Einheiten der Ausgangsdaten abhängen können.

3 Umsetzbarkeit mit dem Mikrozensus

Die Schätzung räumlicher Verteilungen mit dem Mikrozensus Regionalfile 2000 stellt nur eine Machbarkeitsstudie dar. Im Regionalfile liegen die Informationen regional vergrößert auf Ebene der Mikrozensuskreisregionen vor. Außerdem handelt es sich hier um eine Substichprobe aus dem gesamten Mikrozensus 2000 (Christians & Wirth, 2009). Beide Aspekte sollten zu einer erhöhten Varianz der geschätzten räumlichen Verteilungen führen. Daher wurde zunächst versucht, das verwendete Schätzverfahren auf den gesamten Mikrozensus 2015 anzuwenden, wobei eine Ausdehnung auf weitere Mikrozensus-Jahrgänge geplant war. Im Mikrozensus ist es im Onsite-Zugang oder per kontrollierter Datenfernverarbeitung prinzipiell möglich, die Daten auf Gemeindeebene zu analysieren, da in den Onsite-Daten die Gemeindekennziffern vorliegen. Die Shapefiles für die Gemeinden können ohne Weiteres beim Bundesamt für Kartographie und Geodäsie bis zum Jahr 1997 heruntergeladen werden.⁸ Für den ersten Versuch der Übertragung des Verfahrens auf den gesamten Mikrozensus wurde zunächst dieselbe Variablenliste wie mit dem Regionalfile verwendet. Beim Onsite-Zugang und beim Zugang per kontrollierter Datenfernverarbeitung muss gewährleistet sein, dass jede Ergebniszahl auf mindestens drei Fällen beruht (Information und Technik Nordrhein-Westfalen, 2017, S. 17). Da die Erfüllung dieser Auflage bei Ergebnissen aus Kerndichteschätzern unklar ist, wurde zunächst geplant, zuerst Ergebnisse zu produzieren, und nach Erfolg über die Exportbedingungen zu verhandeln.

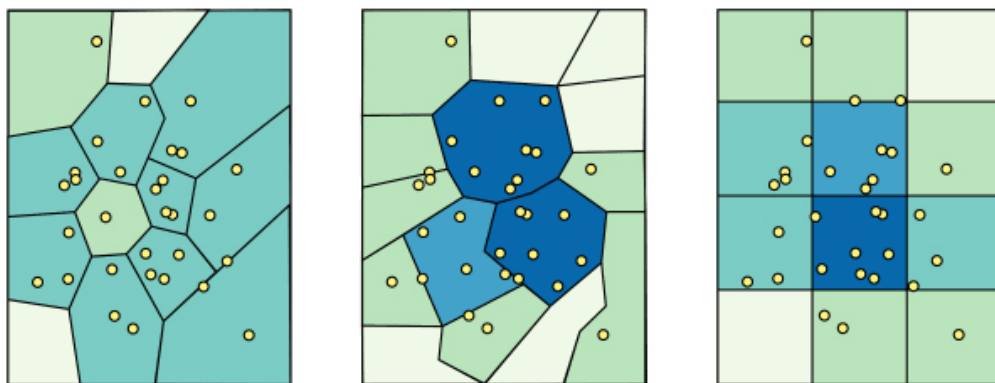
Im Verlauf der Umsetzung mit dem vollen Mikrozensus am Onsite-Arbeitsplatz stellte sich der Plan als nur beschränkt umsetzbar heraus. Erstens stellt das statistische Landesamt Bayern die Gemeindekennziffern nur anonymisiert zur Verfügung, wodurch die Daten in Bayern nicht zu verwenden sind. Zweitens sind die Veränderungen der Gemeindekennziffern, die sich durch Gebietsreformen ergeben, nicht nachvollziehbar. D.h. die Shapefiles aus dem Bundesamt für Kartographie und Geodäsie enthalten zwar die Gemeinden mit Gemeindekennziffer zu klar definierten Stichtagen, diese finden aber keine nachvollziehbare Entsprechung mit den Gemeindekennziffern in den Mikrozensus-Daten. Das Problem ist zwar grundsätzlich lösbar, aber auf Nachfrage beim Forschungsdatenzentrum sehr schwierig. Ohne weiteres waren im Jahr 2015 nur fünf Bundesländer analysierbar, in denen keine Matchprobleme auftraten. Drittens ergibt sich aus dem geschichteten Stichprobenverfahren, dass im Mikrozensus in jedem einzelnen Jahr nicht jede Gemeinde besetzt ist, sondern nur jeder Kreis. Dadurch ist das Kerndichteschätzer-Verfahren von Groß et al. (2017) nicht ohne Weiteres anwendbar. Grundsätzlich ist es möglich, die Informationen der fehlenden Gemeinden durch Small-Area-Schätzer (Fay & Herriot, 1979) oder Multiple Imputation (Buuren, 2018; Rubin, 1987) zu ersetzen. Ohne echte zusätzliche Informationen über die fehlenden Gemeinden ist die Schätzung der räumlichen Verteilungen bei fehlenden Gemeinden nur mit Ausgangsdaten auf Kreisebene möglich bzw. äquivalent dazu. Dieses Problem tritt nur in Saarland, dem kleinsten Flächenland, nicht auf.

Als Konsequenz dieser beiden Probleme wurde das Vorhaben schließlich eingestellt. Erstens kann das Problem der Gemeindeänderung zwar grundsätzlich, aber nur mit hohem Zeitaufwand – insbesondere dadurch, dass die notwendige Einzelfallinspektion in der Onsite-Analyse nur schwierig vorbereitet werden kann – gelöst werden. Zweitens kann das Problem der unvollständigen Abdeckung der Gemeinden nur mit externen validen Hilfsvariablen mit hoher räumlicher Autokorrelation und hoher Korrelation mit anderen Variablen im Mikrozensus auf Gemeindeebene behoben werden..

⁸ Die archivierten Versionen der hier verwendeten Shapefiles sind hier zu finden: https://daten.gdz.bkg.bund.de/produkte/vg/vg250_ebenen_1231/.

4 Modifiable Area Unit Problem im Mikrozensus

Wie bereits mehrfach angesprochen, tritt bei Schätzungen von räumlichen Verteilungen, aber auch bei anderen Analysen von räumlichen Daten das Problem auf, dass die Ergebnisse wesentlich vom gegebenen Messniveau der Ausgangsdaten abhängen. Wenn die Ausgangsdaten stark räumlich aggregiert sind, gehen Unterschiede auf kleinräumigerer Ebene durch die Aggregation verloren. Darüber hinaus hängt die gemessene räumliche Variation von der Form der Aggregationseinheiten ab. Zur Veranschaulichung betrachte man die in Abbildung 10 dargestellten drei räumlichen Aggregationen einer Verteilung von Punkten in der Fläche.



Quelle: Ervin (2012)

Abbildung 10 Abhängigkeit der geschätzten Bevölkerungsdichte von der Form der Aggregationseinheiten

Es ist zu erkennen, dass abhängig von der gewählten Form der räumlichen Aggregationseinheiten die geschätzte Dichte sehr klein (links) oder sehr hoch (Mitte) sein kann.

Der erste Teilaspekt des Problems wird üblicherweise dadurch gelöst, dass man möglichst kleine räumliche Ausgangseinheiten betrachtet. Für den zweiten Aspekt gibt es keine generelle optimale Lösung. Im Folgenden soll untersucht werden, ob die Aggregationsniveaus Gemeinde und Kreis, die man je nach Qualität der Hilfsvariablen und Aufwand bestenfalls im Mikrozensus gegeben hat, ausreichen, um die kleinräumige Variation verschiedener Variablen abzubilden. Dazu soll die Variation verschiedener vergleichbarer Variablen im Zensus 2011 (Statistische Ämter des Bundes und der Länder, 2020) auf Bundesland-, Kreis-, Gemeinde- und 100m-Grid-Ebene und im Mikrozensus SUF 2011 auf Auswahlbezirksebene (siehe FDZ GML, 2020, Durchführung) nebeneinandergestellt werden. Aus den im Zensus und im Mikrozensus gegebenen Variablen lassen sich die Variablen Geschlecht, Familienstand, Alter, Migrationsstatus⁹, Familienform und Haushaltsgröße uneingeschränkt vergleichen. Als Indikator der Variation werden die Segregationsmaße Entropy-Index, Dissimilarity und Isolation (Reardon & O'Sullivan, 2004) verwendet. In Abbildung 11 sind die Segregationsmaße über die räumlichen Einheiten hinweg für die betrachteten Variablen dargestellt

.

⁹ Hier wird der Anteil der nicht in Deutschland geborenen Personen betrachtet.

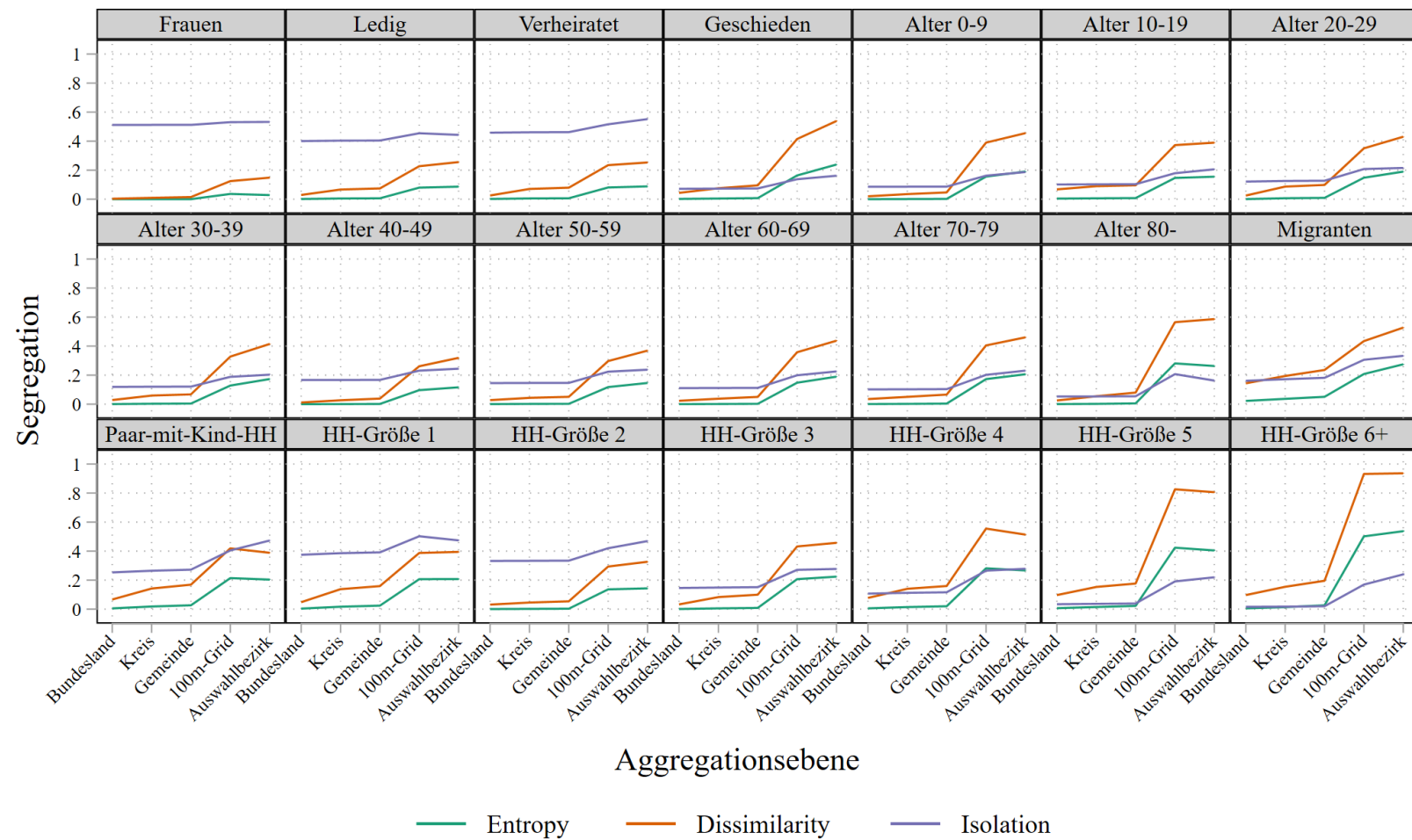


Abbildung 11 Segregation verschiedener soziodemographischer Variablen auf verschiedene räumlichen Aggregationsniveaus

Die Abbildung zeigt auf der X-Achse die räumlichen Aggregationsniveaus Bundesland, Kreis, Gemeinde, 100m-Grid und Auswahlbezirk. Auf der Y-Achse ist das Segregationsausmaß mit den drei Kennzahlen der betrachteten Variablen abgetragen. Alle drei Maße haben einen Wertebereich von 0 bis 1. Es ist zu beachten, dass die Anteile auf Bundesland-, Kreis-, Gemeinde- und 100m-Grid-Ebene aus dem Zensus stammen, und damit Ergebnisse einer Vollerhebung sind. Damit liegt hier kein Stichprobenfehler vor. Bei der Variation der Anteile auf Auswahlbezirksebene stammen die Daten aus dem Mikrozensus SUF 2011, d.h. hier handelt es sich um Schätzer aus einer Stichprobe. Um den Stichprobenfehler abzubilden, wurden Bootstrap Samples gezogen und damit die Variation der Kennzahlen über die Bootstrap Samples geschätzt. Die implizierten 95%-Konfidenzintervalle sind zwar eingezeichnet, aber so klein, dass sie nicht erkennbar sind.

Es zeigt sich nun, dass bei allen betrachteten Variablen die Variation von Bundesland zu Gemeindeebene nur unwesentlich ansteigt. Erst ab der 100m-Grid-Ebene steigt die gemessene räumliche Variation erkennbar an. Der Anstieg zwischen der Gemeindeebene und der 100-Grid-Ebene ist besonders groß beim Anteil der Haushalte mit mehr als 4 Personen. Ein hoher Anstieg zeigt sich aber auch beim Anteil der Geschiedenen, bei fast allen Altersgruppen, beim Migrantenanteil, beim Anteil der Paare, die mit mindestens einem Kind zusammenleben (im Verhältnis zu anderen Formen des Zusammenlebens in einem Haushalt), und bei den kleineren Haushaltsgrößen. Dagegen zeigt sich nur ein schwacher Anstieg beim Frauenanteil und beim Anteil der Ledigen und Verheirateten. Beim Übergang zwischen dem 100m-Grid und der Auswahlbezirksebene, die 6-12 räumlich nah zusammenliegende Haushalte umfasst, findet man bei den meisten Variablen keine Änderung mehr. Ein Anstieg der Dissimilarity um über 5 Punkte findet man beim Anteil der Geschiedenen und beim Migrantenanteil.

Inhaltlich bedeutet dies, dass die deutsche Bevölkerung im Jahr 2011 nicht nach Geschlecht getrennt verteilt lebt. Räumliche Segregation unterhalb der Ebene der Gemeinde zeigt sich hinsichtlich verschiedenster anderer Merkmale. Ethnische Segregation, aber auch Segregation von Geschiedenen lässt sich auch auf kleinräumiger Ebene beobachten, d.h. Migranten und Geschiedene leben kleinräumiger zusammen als bei Aggregationsebene auf 100m-Grid-Ebene erkennbar wäre.

Für die Bewertung der Anwendung des Kerndichteschätzers von Groß et al. (2017) bedeuten diese Ergebnisse, dass der Mikrozensus, mit dem man momentan immer noch bestenfalls auf Gemeindeebene räumliche Verteilungen untersuchen kann¹⁰, nicht ausreichend ist, da die Variation bei fast allen betrachteten Variablen unterhalb der Gemeindeebene wesentlich größer ist als darüber. Diese räumliche Variation geht entsprechend verloren, was dazu führt, dass die Schätzung künstlich geglättete Verteilungen produziert, bzw. kleinräumige Variation nicht aufgelöst wird.

¹⁰ Ab 2018 werden zwar für die befragten Haushalte Geokoordinaten erfasst, diese werden aber gegenwärtig noch nicht veröffentlicht. Siehe hierzu z.B. Gebers und Graze (2019)

5 Zusammenfassung

In diesem Arbeitspapier wurde die Machbarkeitsstudie einer Schätzung von kontinuierlichen räumlichen Verteilungen von soziodemographischen Merkmalen aus dem Mikrozensus dargestellt. Den wesentlichen Teil stellt hier die Umsetzung mit dem Mikrozensus Regionalfile 2000 dar. Mit diesem Datensatz konnten großräumige kontinuierliche Verteilungen mit zufriedenstellendem Stichprobenfehler auf Koordinatenebene geschätzt werden. Da der Regionalfile nur Informationen auf Ebene der Mikrozensuskreisregionen enthält und er nur eine Substichprobe ist, wurde im zweiten Schritt eine Umsetzung des Verfahrens mit dem gesamten Mikrozensus 2000 im Onsite-Zugang versucht. Die Umsetzung stieß auf zwei Probleme. Erstens konnten nur bei fünf Bundesländern ohne aufwändige Aufbereitungsarbeiten die Gemeinde-Shapefiles den Gemeinden im Mikrozensus zugeordnet werden. Zweitens führt das Stichprobenverfahren im Mikrozensus dazu, dass in den meisten Flächenstaaten nicht alle Gemeinden abgedeckt sind. Die fehlenden Gemeinden können zwar grundsätzlich mit Small-Area-Verfahren oder per multipler Imputation aufgefüllt werden, ohne gute Zusatzinformation in Form von zusätzlichen externen Variablen auf Gemeindeebene, kann die Schätzung des Kerndichteschätzers nicht verbessert werden. Diese beiden Hindernisse haben dazu geführt, dass die Umsetzung des Verfahrens mit dem Mikrozensus eingestellt wurde. Abschließend wurde mit den Daten des Zensus 2011 und dem Mikrozensus SUF 2011 untersucht, auf welcher räumlichen Ebene die räumliche Variation verschiedener soziodemographischer Variablen zum Tragen kommt. Daraus ist erkennbar, welche räumliche Aggregationseinheit notwendig ist, um kleinräumige Variation abzubilden. Die Analysen zeigen, dass für die meisten betrachteten Variablen die räumliche Variation unterhalb der Gemeindeebene stattfindet, d.h. eine Aggregation auf Gemeindeebene diese Information verliert. Damit ist der Mikrozensus momentan nicht für die Schätzung räumlicher Kontextdaten geeignet. Dies kann sich grundsätzlich verbessern, wenn die seit 2018 erfassten Geokoordinaten veröffentlicht werden. Da es sich beim Mikrozensus jedoch um eine Klumpenstichprobe von Haushalten aus zufällig gezogenen Auswahlbezirken handelt, ist die zusätzliche Information der genauen Verortung der einzelnen Haushalt möglicherweise kein wesentlicher Zugewinn, da auf Ebene der Haushalte in einem Auswahlbezirk keine räumliche Variation vorliegt.

6 Literatur

- Agostinelli, F., Doepke, M., Sorrenti, G., & Zilibotti, F. (2020). It Takes a Village: The Economics of Parenting with Neighborhood and Peer Effects. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3579206>
- Åslund, O., Blind, I., & Dahlberg, M. (2017). All Aboard? Commuter Train Access and Labor Market Outcomes. *Regional Science and Urban Economics*. <https://doi.org/10.1016/j.regsciurbeco.2017.08.007>
- Behrens, K., Böltken, F., Dittmar, H., Götttsche, F., Gutfleisch, R., Habla, H., Herter-Eschweiler, R., Hoffmann, H., Hoffmeyer-Zlotnik, J. H. P., Klinger, J., Kobl, D., Krack-Roberg, E., Krajzar, H., Krischausky, G., Milbert, A., Mundil-Schwarz, R., Pfister, M., Müller, S., Pavetic, M., ... Wiese, K. (2019). *Regionale Standards: Ausgabe 2019* (Bd. 23). GESIS - Leibniz-Institut für Sozialwissenschaften. <https://doi.org/10.21241/ssoar.62343>
- Bensmann, F., Heling, L., Jünger, S., Mucha, L., Acosta, M., Goebel, J., Meinel, G., Sikder, S., Sure-Vetter, Y., & Zapilko, B. (2020). An Infrastructure for Spatial Linking of Survey Data. *Data Science Journal*, 19(1), 27. <https://doi.org/10.5334/dsj-2020-027>
- Bernasco, W., de Graaff, T., Rouwendal, J., & Steenbeek, W. (2017). Social Interactions and Crime Revisited: An Investigation Using Individual Offender Data in Dutch Neighborhoods. *The Review of Economics and Statistics*, 99(4), 622–636. https://doi.org/10.1162/REST_a_00656
- Best, H., & Rüttenauer, T. (2018). How Selective Migration Shapes Environmental Inequality in Germany: Evidence from Micro-level Panel Data. *European Sociological Review*, 34(1), 52–63. <https://doi.org/10.1093/esr/jcx082>
- Billings, S., Deming, D., & Ross, S. (2016). *Partners in Crime: Schools, Neighborhoods and the Formation of Criminal Networks* (Nr. w21962; S. w21962). National Bureau of Economic Research. <https://doi.org/10.3386/w21962>
- Bischoff, K., & Tach, L. (2020). School Choice, Neighborhood Change, and Racial Imbalance Between Public Elementary Schools and Surrounding Neighborhoods. *Sociological Science*, 7, 75–99. <https://doi.org/10.15195/v7.a4>
- Bluemke, M., Resch, B., Lechner, C., Westerholt, R., & Kolb, J.-P. J.-P. (2017). Integrating Geographic Information into Survey Research: Current Applications, Challenges, and Future Avenues. *Survey Research Methods*, 11(3), 307–327. <https://doi.org/10.18148/srm/2017.v11i3.6733>
- Braga, A. A., Martin Andresen, B. A., Brian Lawton, B., & Braga abraga, A. A. (2017). The Law of Crime Concentration at Places: Editors' Introduction. *Journal of Quantitative Criminology*, 33, 421–426. <https://doi.org/10.1007/s10940-017-9342-0>
- Bruch, E. E., & Mare, R. D. (2006). Neighborhood Choice and Neighborhood Change. *American Journal of Sociology*, 112(3), 667–709. <https://doi.org/10.1086/507856>
- Burdick-Will, J., Grigg, J. A., Nerenberg, K. M., & Connolly, F. (2020). Socially-Structured Mobility Networks and School Segregation Dynamics: The Role of Emergent Consideration Sets. *American Sociological Review*, 0003122420934739. <https://doi.org/10.1177/0003122420934739>
- Buuren, S. van. (2018). *Flexible imputation of missing data* (Second edition). CRC Press, Taylor & Francis Group.
- Chetty, R., & Hendren, N. (2018a). The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects. *The Quarterly Journal of Economics*, 133(3), 1107–1162. <https://doi.org/10.1093/qje/qjy007>

- Chetty, R., & Hendren, N. (2018b). The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates*. *The Quarterly Journal of Economics*, 133(3), 1163–1228. <https://doi.org/10.1093/qje/qjy006>
- Chetty, R., Hendren, N., & Katz, L. F. (2016). The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment. *American Economic Review*, 106(4), 855–902. <https://doi.org/10.1257/aer.20150572>
- Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). Where is the land of Opportunity? The Geography of Intergenerational Mobility in the United States. *The Quarterly Journal of Economics*, 129(4), 1553–1623. <https://doi.org/10.1093/qje/qju022>
- Christians, H., & Wirth, H. (2009). *Mikrozensus-Regionalfile: Hochrechnung, Stichprobenfehler und Anonymisierung*. http://www.forschungsdatenzentrum.de/sites/default/files/mz-2000_regionalfile_doku.pdf
- Clark, W. A. V. (1991). Residential preferences and neighborhood racial segregation: A test of the schelling segregation model. *Demography*, 28(1), 1–19. <https://doi.org/10.2307/2061333>
- Currie, J., Voorheis, J., & Walker, R. (2020). *What Caused Racial Disparities in Particulate Exposure to Fall? New Evidence from the Clean Air Act and Satellite-Based Measures of Air Quality* (Working Paper Nr. 26659). National Bureau of Economic Research. <https://doi.org/10.3386/w26659>
- Dauth, W., & Haller, P. (2020). Is there loss aversion in the trade-off between wages and commuting distances? *Regional Science and Urban Economics*, 103527. <https://doi.org/10.1016/j.regsciurbeco.2020.103527>
- Diekmann, A., & Meyer, R. (2010). Demokratischer Smog? Eine empirische Untersuchung zum Zusammenhang zwischen Sozialschicht und Umweltbelastungen. *Kolner Zeitschrift für Soziologie und Sozialpsychologie*, 62(3), 437–457. <https://doi.org/10.1007/s11577-010-0108-z>
- Dustmann, C., Vasiljeva, K., & Piil Damm, A. (2019). Refugee Migration and Electoral Outcomes. *The Review of Economic Studies*, 86(5), 2035–2091. <https://doi.org/10.1093/restud/rdy047>
- Eckhard, J., & Stauder, J. (2016). Partnermarkt und familiendemografische Prozesse. In J. Stauder, I. Rapp, & J. Eckhard (Hrsg.), *Soziale Bedingungen privater Lebensführung* (S. 121–143). Springer Fachmedien. https://doi.org/10.1007/978-3-658-10986-8_6
- Eckhard, J., & Stauder, J. (2018). Migration and the Partner Market: How Gender-Selective Relocations Affect Regional Mating Chances in Germany. *European Journal of Population*, 34(1), 59–86. <https://doi.org/10.1007/s10680-017-9422-8>
- Ervin, D. (2012). *MAUP – GISPopSci*. MAUP. <http://gispopsci.org/maup/>
- Fay, R. E., & Herriot, R. A. (1979). Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, 74(366), 269–277. <https://doi.org/10.2307/2286322>
- FDZ GML. (2020). *Erhebung: MZ*. <https://www.gesis.org/missy/metadata/MZ/>
- Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder. (2018). *Mikrozensus-Regionalfile 2000, SUF, Version 0* [Data set]. RDC of the Federal Statistical Office and the statistical offices of the Länder (RDC). <https://doi.org/10.21242/12211.2000.00.06.3.1.0>
- Friedrichs, J., & Nonnenmacher, A. (Hrsg.). (2014). *Soziale Kontexte und Soziale Mechanismen*. VS Verlag für Sozialwissenschaften. <https://www.springer.com/de/book/9783658066932>
- Galster, G. C. (2012). The Mechanism(s) of Neighbourhood Effects: Theory, Evidence, and Policy Implications. In M. van Ham, D. Manley, N. Bailey, L. Simpson, & D. Maclennan (Hrsg.), *Neighbourhood Effects Research: New Perspectives* (S. 23–56). Springer Netherlands. https://doi.org/10.1007/978-94-007-2309-2_2

- Gebers, K., & Graze, P. (2019). Statistische Datengewinnung durch die Nutzung geografischer Informationen. *WISTA*, 4, 11–18.
- GeoBasis-DE, & Bundesamt für Kartographie und Geodäsie. (2000). *VG250 (Ebenen) Historische Daten Verwaltungsgebiete der Bundesrepublik Deutschland mit/ohne Einwohnerzahlen*. https://daten.gdz.bkg.bund.de/produkte/vg/vg250_ebenen_1231/2000/vg250_12-31.utm32w.shape.ebenen.zip
- González, T. A., & Granic, G. D. (2020). Spatial Voting Meets Spatial Policy Positions: An Experimental Appraisal. *American Political Science Review*, 114(1), 285–290. <https://doi.org/10.1017/S0003055419000492>
- Groß, M., & Rendtel, U. (2016). Kernel Density Estimation for Heaped Data. *Journal of Survey Statistics and Methodology*, 4(3), 339–361. <https://doi.org/10.1093/jssam/smw011>
- Groß, M., Rendtel, U., Schmid, T., Schmon, S., & Tzavidis, N. (2017). Estimating the density of ethnic minorities and aged people in Berlin: Multivariate kernel density estimation applied to sensitive georeferenced administrative data protected via measurement error. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(1), 161–183. <https://doi.org/10.1111/rssa.12179>
- Halla, M., Wagner, A. F., & Zweimüller, J. (2017). Immigration and Voting for the Far Right. *Journal of the European Economic Association*, 15(6), 1341–1385. <https://doi.org/10.1093/jeea/jvx003>
- Ham, M. van, Manley, D., Bailey, N., Simpson, L., & MacLennan, D. (Hrsg.). (2012). *Neighbourhood Effects Research: New Perspectives*. Springer Netherlands. <https://doi.org/10.1007/978-94-007-2309-2>
- Heissel, J., Persico, C., & Simon, D. (2019). *Does Pollution Drive Achievement? The Effect of Traffic Pollution on Academic Performance*. 64.
- Hermansen, A. S., Borgen, N. T., & Mastekaasa, A. (2020). Long-Term Trends in Adult Socio-Economic Resemblance between Former Schoolmates and Neighbouring Children. *European Sociological Review*, 36(3), 366–380. <https://doi.org/10.1093/esr/jcz066>
- Heuermann, D., & Schmieder, J. (2018). *The Effect of Infrastructure on Worker Mobility: Evidence from High-Speed Rail Expansion in Germany*. National Bureau of Economic Research. <https://doi.org/10.3386/w24507>
- Horr, A. (2016). Nachbarschaftseffekte. In C. Diehl, C. Hunkler, & C. Kristen (Hrsg.), *Ethnische Ungleichheiten im Bildungsverlauf* (S. 397–430). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-04322-3_9
- Information und Technik Nordrhein-Westfalen. (2017). *Regelungen zur Auswertung von Mikrodaten in den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder (FDZ)* (Forschungsdatenzentren der Statistischen Ämter, Hrsg.). https://www.forschungsdatenzentrum.de/sites/default/files/fdz_broschuere_regelungen.pdf
- INSPIRE Thematic Working Group Coordinate Reference Systems & Geographical Grid Systems. (2014). *D2.8.1.1 Data Specification on coordinate Reference Systems – Technical Guidelines*. <https://inspire.ec.europa.eu/id/document/tg/rs>
- Kawabata, M., & Abe, Y. (2017). Intra-metropolitan spatial patterns of female labor force participation and commute times in Tokyo. *Regional Science and Urban Economics*. <https://doi.org/10.1016/j.regsciurbeco.2017.11.003>
- Kohlhuber, M., Mielck, A., Weiland, S. K., & Bolte, G. (2006). Social inequality in perceived environmental exposures in relation to housing conditions in Germany. *Environmental Research*, 101(2), 246–255. <https://doi.org/10.1016/j.envres.2005.09.008>

- Kroll, M., & Schnell, R. (2016). Anonymisation of geographical distance matrices via Lipschitz embedding. *International Journal of Health Geographics*, 15(1), 1. <https://doi.org/10.1186/s12942-015-0031-7>
- Lee, B. A., Reardon, S. F., Firebaugh, G., Farrell, C. R., Matthews, S. A., & O'Sullivan, D. (2008). Beyond the Census Tract: Patterns and Determinants of Racial Segregation at Multiple Geographic Scales. *American Sociological Review*, 73(5), 766–791. <https://doi.org/10.1177/000312240807300504>
- Legewie, J. (2018). Living on the Edge: Neighborhood Boundaries and the Spatial Dynamics of Violent Crime. *Demography*, 55(5), 1957–1977. <https://doi.org/10.1007/s13524-018-0708-1>
- Legewie, J., & Schaeffer, M. (2016). Contested Boundaries: Explaining Where Ethnoracial Diversity Provokes Neighborhood Conflict. *American Journal of Sociology*, 122(1), 125–161. <https://doi.org/10.1086/686942>
- Lewis-Beck, M. S., & Stegmaier, M. (2000). Economic determinants of electoral outcomes. *Annual Review of Political Science*, 3, 183–219.
- Margalit, Y. (2019). Political Responses to Economic Shocks. *Annual Review of Political Science*, 22(1), 277–295. <https://doi.org/10.1146/annurev-polisci-050517-110713>
- Möller, J., & Zierer, M. (2018). Autobahns and jobs: A regional study using historical instrumental variables. *Journal of Urban Economics*, 103, 18–33. <https://doi.org/10.1016/j.jue.2017.10.002>
- Oberwittler, D., Rabold, S., & Baier, D. (Hrsg.). (2013). *Städtische Armutsquartiere--kriminelle Lebenswelten? Studien zu sozialräumlichen Kontexteffekten auf Jugendkriminalität und Kriminalitätswahrnehmungen*. Springer VS.
- Openshaw, S. (1983). *The modifiable areal unit problem*. Geo Books.
- Pestel, N., & Wozny, F. (2019). *Low Emission Zones for Better Health: Evidence from German Hospitals*. 51.
- Reardon, S. F., & Firebaugh, G. (2002). Measures of Multigroup Segregation. *Sociological Methodology*, 32(1), 33–67. <https://doi.org/10.1111/1467-9531.00110>
- Reardon, S. F., & O'Sullivan, D. (2004). Measures of Spatial Segregation. *Sociological Methodology*, 34(1), 121–162. <https://doi.org/10.1111/j.0081-1750.2004.00150.x>
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.
- Schweers, S., Kinder-kurlanda, K., Müller, S., & Siegers, P. (2016). Conceptualizing a Spatial Data Infrastructure for the Social Sciences: An Example from Germany. *Journal of Map & Geography Libraries*, 12(1), 100–126. <https://doi.org/10.1080/15420353.2015.1100152>
- Statistische Ämter des Bundes und der Länder. (2020). *ZENSUS2011—Bevölkerungs- und Wohnungszählung 2011—Ergebnisse des Zensus 2011 zum Download—Erweitert*. <https://www.zensus2011.de/DE/Home/Aktuelles/DemografischeGrunddaten.html?nn=3065474>
- Stauder, J. (2011). Regionale ungleichheit auf dem partnermarkt? Die makrostrukturellen rahmenbedingungen der partnerwahl in regionaler perspektive. *Soziale Welt*, 62(1), 45–73.
- Stauder, J. (2015). Durchdringende Sozialstruktur? Der Einfluss makrostruktureller Rahmenbedingungen auf den Partnermarkt. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 67(3), 401–432. <https://doi.org/10.1007/s11577-015-0340-7>
- Süß, P. (2020). Regional Market Income Inequality and its Impact on Crime in Germany: A Spatial Panel Data Approach with Local Spillovers. *Jahrbücher Für Nationalökonomie Und Statistik*, 240(4), 387–415. <https://doi.org/10.1515/jbnst-2018-0052>

- Vom Berge, P., & Schmillen, A. (2015). Direct and indirect effects of mass layoffs: Evidence from geo-referenced data. In *IAB Discussion Paper* (Nr. 201511; IAB Discussion Paper). Institut für Arbeitsmarkt- und Berufsforschung (IAB), Nürnberg [Institute for Employment Research, Nuremberg, Germany]. <https://ideas.repec.org/p/iab/iabdpa/201511.html>
- Weßling, K. D. (2016). *The influence of socio-spatial contexts on transitions from school to vocational and academic training in Germany*.
- Wickes, R., & Hipp, J. R. (2018). The Spatial and Temporal Dynamics of Neighborhood Informal Social Control and Crime. *Social Forces*, 97(1), 277–308. <https://doi.org/10.1093/sf/soy026>

7 Appendix

Code 1 Aufbereitung der Aggregatdaten aus dem Mikrozensus Regionalfile 2000

```

* Estimation of regional Means and Covariance Matrices of /*
*/ specific variables with MZ REgionalfile 2000
* Variables
* Geographical unit: MZKR (Mikrozensuskreisregion)
* Substantive Variables:
* Schulabschluss
* Schulbesuch Oberstufe in Alter 18
* Travel time to school
* Household income decile

* Haushaltsvariable ef627
* Varianz Zufallszahl

clear *
set more off
use "R:\mzr\1506-2010_Regionalfile_MZ_2000\STATA\mz00_regiona
lfile.dta", clear
keep ef711 /*mzkr*/ ef286 ef287 /*schulabschluss*/ ef71 ef72 ef30 /*
*/ /*schulbesuch*/ ef1 ef710 ef3s ef4s ef5s ef28 ef643 ef506 /*
*/ ef507 ef39 ef32 /* homo*/ ef83 /* weg zur schule
*/ ef539 /* haushaltsnetto */ ef43 ef44

* Nationalität
gen national=1 if ef43==1
replace national=2 if ef44==23
replace national=3 if ef44==10
replace national=4 if inlist(ef44,20,11,2)
replace national=5 if ef44==8
replace national=6 if ef44==6
replace national=7 if ef44==15
replace national=8 if national==.
label def national 1 "1 deutsch" 2 "2 tuerk" 3 "3 ital" 4 /*
*/ "4 ex-jugo" 5 "5 gus" 6 "6 griechenland" 7 "7 polen" 8 /*
*/ "8 sonst"
label val national national
label var national "1. stabue"

* Schulabschluss
tab ef286 ef287, m
gen schulabschl=1 if ef286==8 | ef287==1
replace schulabschl=2 if ef287==2 | ef287==3
replace schulabschl=3 if ef287==4 | ef287==5
label def schulabschl 1 "1 kein/haupt" 2 "2 real/pos" 3 /*
*/ "3 abi/fachabi"
label val schulabschl schulabschl
label var schulabschl "hoechster schulabschl (ohne schueler)"

* In gymnasialer Oberstufe im Alter von 18-19
gen oberstufe=ef72==3 if ef30==18
label var oberstufe "Klasse 11-13 im Alter von 18"

* soziale mobilität,
bysort ef506 ef1 ef710 ef3s ef4s ef5s : gen npers=_N if ef506==1
bysort npers ef506 ef1 ef710 ef3s ef4s (ef5s): /*
*/ egen exactly_one_main_person=total(ef507==1) if /*
*/ ef506==1 & npers==1
gen kind=ef507==3 if exactly_one_main_person==1 & /*
*/ ef506==1 & npers==1
gen partner=ef507==2 | ef39==1 if exactly_one_main_person==1 /*
*/ & ef506==1 & npers==1
bysort exactly_one_main_person npers ef506 ef1 ef710 ef3s /*
*/ ef4s (ef5s): egen npartner=total(partner==1) if /*
*/ exactly_one_main_person==1 & ef506==1 & npers==1
gen ego=ef507==1 if exactly_one_main_person==1 & ef506==1 & /*
*/ npers==1

```

```

gen egoabi=schulabschl==3 if ego==1
gen partnerabi=schulabschl==3 if partner==1
bysort exactly_one_main_person npers ef506 ef1 ef710 ef3s ef4s /*
  */ (ef5s): egen one_parent_with_abi=total(egoabi==1 | /*
  */ partnerabi==1) if exactly_one_main_person==1 & ef506==1 & /*
  */ npers==1
replace one_parent_with_abi=1 if one_parent_with_abi==2
gen mobility=1 if oberstufe==0 & one_parent_with_abi==1 & kind==1
replace mobility=2 if (oberstufe==0 & one_parent_with_abi==0) | /*
  */ (oberstufe==1 & one_parent_with_abi==1) & kind==1
replace mobility=3 if oberstufe==1 & one_parent_with_abi==0 & /*
  */ kind==1
label var mobility "Bildungsmobilität"
label def mobility 1 "1 Kind hat kein Abi, mind. 1 Elternteil hat /*
  */ Abi" 2 "2 Kind und Eltern habe beide Abi oder beide nicht" 3 /*
  */ "3 Kind hat Abi, Eltern nicht"
label val mobility mobility
drop npers exactly_one_main_person npartner one_parent_with_abi /*
  */ kind ego partner egoabi partnerabi

* Gleichgeschlechtliche kohabtierende Partnerschaft
bysort ef506 ef1 ef710 ef3s ef4s ef5s : gen npers=_N if ef506==1
bysort npers ef506 ef1 ef710 ef3s ef4s (ef5s): egen /*
  */ exactly_one_main_person=total(ef507==1) if ef506==1 & npers==1
gen partner=ef507==2 | ef39==1 if exactly_one_main_person==1 & /*
  */ ef506==1 & npers==1
bysort exactly_one_main_person npers ef506 ef1 ef710 ef3s ef4s /*
  */ (ef5s): egen npartner=total(partner==1) if /*
  */ exactly_one_main_person==1 & ef506==1 & npers==1
bysort npartner exactly_one_main_person npers ef506 ef1 ef710 /*
  */ ef3s ef4s (ef5s): egen nmale=total(ef32==1) if npartner==1 & /*
  */ exactly_one_main_person==1 & ef506==1 & npers==1 & (ef507==1 /*
  */ | (ef507==2 | (ef507!=1 & ef39==1)))
bysort npartner exactly_one_main_person npers ef506 ef1 ef710 /*
  */ ef3s ef4s (ef5s): egen nfemale=total(ef32==2) if npartner==1 /*
  */ & exactly_one_main_person==1 & ef506==1 & npers==1 & /*
  */ (ef507==1 | (ef507==2 | (ef507!=1 & ef39==1)))
gen homo=nfemale==2 | nmale==2 if nfemale!=. & nmale!=.
* gen homo=inlist(ef630,5,6,7) if ef629!=.
label var homo "Gleichgeschlechtlich kohabitierend"
label def homo 1 "1 gleichgeschl" 0 "0 gegengeschl"
gen homo_male=nmale==2 if nfemale!=. & nmale!=.
label var homo_male "Männliche gleichgeschlechtlich kohabitierende"
gen homo_female=nfemale==2 if nfemale!=. & nmale!=.
label var homo_female "Männliche gleichgeschlechtlich /*
  */ kohabitierende"

* Travel time to school
replace ef83=. if ef83>4
rename ef83 schulweg

* Haushaltsnetto
capture drop npers
bysort ef506 ef1 ef710 ef3s ef4s ef5s : gen npers=_N if ef506==1
bysort npers ef506 ef1 ef710 ef3s ef4s (ef5s): egen check_var=sd(ef539) if ef506==1 &
npers==1
bysort npers ef506 ef1 ef710 ef3s ef4s (ef5s): replace ef539=. if /*
  */ _n!=1 & ef506==1 & npers==1 & ef539<25
gen hheinkkat=1 if inrange(ef539,1,6)
replace hheinkkat=2 if inrange(ef539,7,9)
replace hheinkkat=3 if inrange(ef539,10,12)
replace hheinkkat=4 if inrange(ef539,13,24)
label var hheinkkat "HH Nettoinkommen im April (ohne /*
  */ Selbstständige), Quartile"
label def hheinkkat 1 "1 Bis unter 2200 DM " 2 "2 2200 bis unter /*
  */ 3500 DM" 3 "3 3500 bis unter 5000 DM" 4 "4 5000 DM und mehr"
label val hheinkkat hheinkkat

* Zufallsvariable
set seed 1
gen random=runiformint(0,1)
label var random "coin flip"

```

```

* keep relevant variables
keep ef711 national schulabschl oberstufe mobility homo schulweg /*
  */ hheinkkat
order ef711 national schulabschl oberstufe mobility homo schulweg /*
  */ hheinkkat
compress

* generate proper ID
* gen SHN=string(ef711,"%05.0f")+ "00000"
gen double SHN=ef711*100000
order SHN

* create aggregate file
quietly ds national schulabschl oberstufe mobility homo schulweg /*
  */ hheinkkat

/*
gen ding=ef83!=.
collapse (count) ding, by(SHN ef83)
drop if ef83==.
rename ef83 values
rename ding ef83_N
bysort SHN: egen elig_N=total(ef83_N)

reshape i SHN
reshape j values
reshape xij ef83_N elig_N
reshape wide

if `first'==0 {
  merge 1:1 SHN using "P:\MZKR-GEO\mzr2000_agg.dta", nogen /*
    */ noreport
    save "P:\MZKR-GEO\mzr2000_agg.dta", replace
}
if `first'==1 {
  save "P:\MZKR-GEO\mzr2000_agg.dta", replace
  local first=0
}
*/

local first=1
foreach var in `r(varlist)' {
  preserve
  gen ding=`var'!=.
  collapse (count) ding, by(SHN `var')
  drop if `var'==.
  rename `var' values
  rename ding `var'_N
  bysort SHN: egen `var'_NE=total(`var'_N)

  reshape i SHN
  reshape j values
  reshape xij `var'_N `var'_NE
  reshape wide

  quietly ds *_N*
  local list1=`r(varlist)'"
  quietly ds *_NE*, not
  local list2=`r(varlist)'"
  local list3:list list1 - list2
  egen max=rowmax(`list3')
  foreach var in `list3' {
    replace `var'=max
  }
  drop max

  if `first'==0 {
    merge 1:1 SHN using "P:\MZKR-GEO\v2\mzr2000_agg_v5.dta", /*
      */ nogen noreport
      save "P:\MZKR-GEO\v2\mzr2000_agg_v5.dta", replace
  }
}

```

```
if `first'==1 {
    save "P:\MZKR-GEO\v2\mzr2000_agg_v5.dta", replace
    local first=0
}
restore
}
use "P:\MZKR-GEO\v2\mzr2000_agg_v5.dta", clear
quietly ds SHN, not
foreach var in `r(varlist)' {
    recode `var' .=0
}
compress
sort SHN
save "P:\MZKR-GEO\v2\mzr2000_agg_v5.dta", replace

exit
```

Code 2 R-Code zur Erstellung der Mikrozensuskreisregion-Shapes und der Schätzung der kontinuierlichen räumlichen Verteilungen

```
require(sp, maptools, rgdal, readstata13)

rm(list = ls())
setwd("P:/MZKR-GEO")
set.seed(1)
# Import Kreis-Shapes
kreise<-rgdal::readOGR("P:/MZKR-GEO/Verwaltungsgebietelzu250000/vg250_2000-01-
01.utm32w.shape.ebenen/vg250_ebenen-historisch/de0001", "vg250krs")
# Transform Kreis-Shapes to area-preserving projection, as Equal Area Grids based on
ETRS89-LAEA (EPSG 3035)
kreise<-sp::spTransform(kreise, sp::CRS("+init=epsg:3035"))

# Drop Bodensee
kreise<-kreise[kreise$SHN!="0800000000"&kreise$SHN!="0900000000",]

# Reproduce fusion of Kreise as was done in the Mikrozensus Regionalfile 2000
# following Documentation on DVD and
# Helga Christians / Heike Wirth
# Mikrozensus-Regionalfile:
#   Hochrechnung, Stichprobenfehler und Anonymisierung
#   (unter Mitarbeit von:
#     Baris Bayindir und Reinhard Schunck)
#   Düsseldorf, Mannheim
#   November 2009
kreise$SHN[kreise$SHN=="0100400000"]<-"0105800000"
kreise$SHN[kreise$SHN=="0100100000"]<-"0105900000"
kreise$SHN[kreise$SHN=="0315600000"]<-"0315500000"
kreise$SHN[kreise$SHN=="0325500000"]<-"0315500000"
kreise$SHN[kreise$SHN=="0340100000"]<-"0325100000"
kreise$SHN[kreise$SHN=="0346100000"]<-"0335200000"
kreise$SHN[kreise$SHN=="0336000000"]<-"0335400000"
kreise$SHN[kreise$SHN=="0340200000"]<-"0345200000"
kreise$SHN[kreise$SHN=="0346200000"]<-"0345500000"
kreise$SHN[kreise$SHN=="0340500000"]<-"0345500000"
kreise$SHN[kreise$SHN=="0643700000"]<-"0643100000"
kreise$SHN[kreise$SHN=="0713400000"]<-"0713300000"
kreise$SHN[kreise$SHN=="0713500000"]<-"0714000000"
kreise$SHN[kreise$SHN=="0723200000"]<-"0723300000"
kreise$SHN[kreise$SHN=="0721100000"]<-"0723500000"
kreise$SHN[kreise$SHN=="0731900000"]<-"0733100000"
kreise$SHN[kreise$SHN=="0731600000"]<-"0733200000"
kreise$SHN[kreise$SHN=="0733600000"]<-"0733300000"
kreise$SHN[kreise$SHN=="0731300000"]<-"0733700000"
kreise$SHN[kreise$SHN=="0731100000"]<-"0733800000"
kreise$SHN[kreise$SHN=="0731800000"]<-"0733800000"
kreise$SHN[kreise$SHN=="0731700000"]<-"0734000000"
kreise$SHN[kreise$SHN=="0732000000"]<-"0734000000"
kreise$SHN[kreise$SHN=="0821100000"]<-"0821600000"
kreise$SHN[kreise$SHN=="0918200000"]<-"0917300000"
kreise$SHN[kreise$SHN=="0918500000"]<-"0918600000"
kreise$SHN[kreise$SHN=="0916300000"]<-"0918700000"
kreise$SHN[kreise$SHN=="0917200000"]<-"0918900000"
kreise$SHN[kreise$SHN=="0918000000"]<-"0919000000"
kreise$SHN[kreise$SHN=="0927600000"]<-"0927200000"
kreise$SHN[kreise$SHN=="0926100000"]<-"0927400000"
kreise$SHN[kreise$SHN=="0926200000"]<-"0927500000"
kreise$SHN[kreise$SHN=="0927900000"]<-"0927700000"
kreise$SHN[kreise$SHN=="0926300000"]<-"0927800000"
kreise$SHN[kreise$SHN=="0936100000"]<-"0937100000"
kreise$SHN[kreise$SHN=="0936300000"]<-"0937400000"
kreise$SHN[kreise$SHN=="0937700000"]<-"0937400000"
kreise$SHN[kreise$SHN=="0946100000"]<-"0947100000"
kreise$SHN[kreise$SHN=="0946200000"]<-"0947200000"
kreise$SHN[kreise$SHN=="0947700000"]<-"0947200000"
kreise$SHN[kreise$SHN=="0946300000"]<-"0947300000"
kreise$SHN[kreise$SHN=="0946400000"]<-"0947500000"
kreise$SHN[kreise$SHN=="0947900000"]<-"0947500000"
kreise$SHN[kreise$SHN=="0947600000"]<-"0947800000"
```

```

kreise$SHN[kreise$SHN=="0956100000"]<-"0957100000"
kreise$SHN[kreise$SHN=="0957700000"]<-"0957100000"
kreise$SHN[kreise$SHN=="0957500000"]<-"0957100000"
kreise$SHN[kreise$SHN=="0956500000"]<-"0957600000"
kreise$SHN[kreise$SHN=="0966100000"]<-"0967100000"
kreise$SHN[kreise$SHN=="0967300000"]<-"0967400000"
kreise$SHN[kreise$SHN=="0966200000"]<-"0967800000"
kreise$SHN[kreise$SHN=="0967500000"]<-"0967900000"
kreise$SHN[kreise$SHN=="0976200000"]<-"0977700000"
kreise$SHN[kreise$SHN=="0976400000"]<-"0977800000"
kreise$SHN[kreise$SHN=="0977300000"]<-"0977900000"
kreise$SHN[kreise$SHN=="0976300000"]<-"0978000000"
kreise$SHN[kreise$SHN=="0977600000"]<-"0978000000"
kreise$SHN[kreise$SHN=="1004600000"]<-"1004200000"
kreise$SHN[kreise$SHN=="1205300000"]<-"1206700000"
kreise$SHN[kreise$SHN=="1207000000"]<-"1206800000"
kreise$SHN[kreise$SHN=="1205100000"]<-"1206900000"
kreise$SHN[kreise$SHN=="1300200000"]<-"1305500000"
kreise$SHN[kreise$SHN=="1305200000"]<-"1305600000"
kreise$SHN[kreise$SHN=="1300500000"]<-"1305700000"
kreise$SHN[kreise$SHN=="1306100000"]<-"1305700000"
kreise$SHN[kreise$SHN=="1300600000"]<-"1305800000"
kreise$SHN[kreise$SHN=="1300100000"]<-"1305900000"
kreise$SHN[kreise$SHN=="1306200000"]<-"1305900000"
kreise$SHN[kreise$SHN=="1418800000"]<-"1417300000"
kreise$SHN[kreise$SHN=="1416600000"]<-"1417800000"
kreise$SHN[kreise$SHN=="1417100000"]<-"1418100000"
kreise$SHN[kreise$SHN=="1426300000"]<-"1428400000"
kreise$SHN[kreise$SHN=="1426400000"]<-"1429200000"
kreise$SHN[kreise$SHN=="1437500000"]<-"1438900000"
kreise$SHN[kreise$SHN=="1510100000"]<-"1515400000"
kreise$SHN[kreise$SHN=="1515300000"]<-"1515900000"
kreise$SHN[kreise$SHN=="1515100000"]<-"1517100000"
kreise$SHN[kreise$SHN=="1526600000"]<-"1526000000"
kreise$SHN[kreise$SHN=="1526500000"]<-"1526100000"
kreise$SHN[kreise$SHN=="1526800000"]<-"1526100000"
kreise$SHN[kreise$SHN=="1536400000"]<-"1535200000"
kreise$SHN[kreise$SHN=="1536700000"]<-"1535200000"
kreise$SHN[kreise$SHN=="1536200000"]<-"1535500000"
kreise$SHN[kreise$SHN=="1536900000"]<-"1535700000"
kreise$SHN[kreise$SHN=="1606500000"]<-"1606200000"
kreise$SHN[kreise$SHN=="1605600000"]<-"1606300000"
kreise$SHN[kreise$SHN=="1605400000"]<-"1606600000"
kreise$SHN[kreise$SHN=="1606800000"]<-"1606700000"
kreise$SHN[kreise$SHN=="1605500000"]<-"1607100000"
kreise$SHN[kreise$SHN=="1606900000"]<-"1607200000"
kreise$SHN[kreise$SHN=="1605300000"]<-"1607400000"
# Sort Kreise in the order of the serial number SHN
kreise<-kreise[order(kreise$SHN),]
# Combine Kreise that have to be fused because of anonymity
mzkr<-maptools::unionSpatialPolygons(kreise,kreise$SHN)

# Merge MZ-Regionalfile 2000
# import stata file
mzr2000_agg<-readstata13::read.dta13("P:/MZKR-GEO/v2/mzr2000_agg_v5.dta",
                                     con=
vert.factors=T,encoding="latin1",missing.type=T)

# put label points and shape ids from MZKR polygons in data frame
data<-
as.data.frame(cbind(sp::coordinates(mzkr),as.numeric(sp::getSpPPolygonsIDSlots(mzkr))
))
names(data)[names(data)=="V1"]<-"x"
names(data)[names(data)=="V2"]<-"y"
names(data)[names(data)=="V3"]<-"SHN"

mzr2000_agg<-merge.data.frame(data,mzr2000_agg,by.x="SHN",by.y="SHN",all.x=T,all.y=T)
rm(data)
mzr2000_agg<-mzr2000_agg[order(mzr2000_agg$SHN),]

# Set parameters for kernel heaping
b<-10 # number of burnin samples

```

```

s<-30 # number of estimatione samples
gs<-500 # grid size
nc<-64 # number of clusters for parallelization
nt<-8 # number of threads for parallelization

#####
# Estimate distributions
#####
#-----
# HH income quartile 1
#-----
# Estimate using columns 2 and 3 as coordinates of centroids of MZ Kreisregionen
# and columns 4 as number of household in first quartile
# and columns 5 as number of household at all
# shapefiles are in object mzk
est<-Kernelheaping::dshapebivrProp(mzr2000_agg[,c(2:3,4:5)],burnin=b, samples=s,
adaptive=F,
                                shapefile=mzkr, gridsize=gs,
                                boundary=F,deleteShapes =
NULL,fastWeights=F,numChains = nc,numThreads = nt)
# init pdf document for output
pdf("P:/MZKR-GEO/v2/hhnettoeinkq1.pdf")
# plot estimated distribution z over coordinates x and y (implied by grid size and
boundary of mzk)
fields::image.plot(x=est$Mestimates$eval.points[[1]],
                   y=est$Mestimates$eval.points[[2]],
                   z=est$proportions,
                   col=RColorBrewer::brewer.pal(10,"RdYlGn"),
                   breaks=(0:10)/10,lab.breaks=(0:10)/10)
# overlay with original map of Kreise
sp::plot(kreise,add=T)
title("Anteil HH-NettoEink 1.Quartil")
# save
dev.off()
#-----
# HH income quartile 2
#-----
# Estimate using columns 2 and 3 as coordinates of centroids of MZ Kreisregionen
# and columns 6 as number of household in 2nd quartile
# and columns 7 as number of household at all
# shapefiles are in object mzk
est<-Kernelheaping::dshapebivrProp(mzr2000_agg[,c(2:3,6:7)],burnin=b, samples=s,
adaptive=F,
                                shapefile=mzkr, gridsize=gs,
                                boundary=F,deleteShapes =
NULL,fastWeights=F,numChains = nc,numThreads = nt)
pdf("P:/MZKR-GEO/v2/hhnettoeinkq2.pdf")
fields::image.plot(x=est$Mestimates$eval.points[[1]],
                   y=est$Mestimates$eval.points[[2]],
                   z=est$proportions,
                   col=RColorBrewer::brewer.pal(10,"RdYlGn"),
                   breaks=(0:10)/10,lab.breaks=(0:10)/10)
sp::plot(kreise,add=T)
title("Anteil HH-NettoEink 2.Quartil")
dev.off()
#-----
# HH income quartile 3
#-----
# Estimate using columns 2 and 3 as coordinates of centroids of MZ Kreisregionen
# and columns 8 as number of household in 3nd quartile
# and columns 9 as number of household at all
# shapefiles are in object mzk
est<-Kernelheaping::dshapebivrProp(mzr2000_agg[,c(2:3,8:9)],burnin=b, samples=s,
adaptive=F,
                                shapefile=mzkr, gridsize=gs,
                                boundary=F,deleteShapes =
NULL,fastWeights=F,numChains = nc,numThreads = nt)
pdf("P:/MZKR-GEO/v2/hhnettoeinkq3.pdf")
fields::image.plot(x=est$Mestimates$eval.points[[1]],
                   y=est$Mestimates$eval.points[[2]],
                   z=est$proportions,
                   col=RColorBrewer::brewer.pal(10,"RdYlGn"),

```

```

                                breaks=(0:10)/10,lab.breaks=(0:10)/10)
sp::plot(kreise,add=T)
title("Anteil HH-NettoEink 3.Quartil")
dev.off()
#-----
# HH income quartile 4
#-----
# Estimate using columns 2 and 3 as coordinates of centroids of MZ Kreisregionen
# and columns 10 as number of household in 4th quartile
# and columns 11 as number of household at all
# shapefiles are in object mzkr
est<-Kernelheaping::dshapebivrProp(mzr2000_agg[,c(2:3,10:11)],burnin=b, samples=s,
adaptive=F,
                                shapefile=mzkr, gridsize=gs,
                                boundary=F,deleteShapes =
NULL,fastWeights=F,numChains = nc,numThreads = nt)
pdf("P:/MZKR-GEO/v2/hhnettoeinkq4.pdf")
fields::image.plot(x=est$Mestimates$eval.points[[1]],
                   y=est$Mestimates$eval.points[[2]],
                   z=est$proportions,
                   col=RColorBrewer::brewer.pal(10,"RdYlGn"),
                   breaks=(0:10)/10,lab.breaks=(0:10)/10)
sp::plot(kreise,add=T)
title("Anteil HH-NettoEink 4.Quartil")
dev.off()
#-----
# way to school less than 10 minutes
#-----
# Estimate using columns 2 and 3 as coordinates of centroids of MZ Kreisregionen
# and columns 12 as number of children with respective length of commute to scholl
# and columns 13 as number of all children
# shapefiles are in object mzkr
est<-Kernelheaping::dshapebivrProp(mzr2000_agg[,c(2:3,12:13)],burnin=b, samples=s,
adaptive=F,
                                shapefile=mzkr, gridsize=gs,
                                boundary=F,deleteShapes =
NULL,fastWeights=F,numChains = nc,numThreads = nt)
pdf("P:/MZKR-GEO/v2/schulweg_lt10min.pdf")
fields::image.plot(x=est$Mestimates$eval.points[[1]],
                   y=est$Mestimates$eval.points[[2]],
                   z=est$proportions,
                   col=RColorBrewer::brewer.pal(10,"RdYlGn"),
                   breaks=(0:10)/10,lab.breaks=(0:10)/10)
sp::plot(kreise,add=T)
title("Anteil Schulweg unter 10 min")
dev.off()
#-----
# way to school 10 minutes or more and less than 30 minutes
#-----
# Estimate using columns 2 and 3 as coordinates of centroids of MZ Kreisregionen
# and columns 14 as number of children with respective length of commute to scholl
# and columns 15 as number of all children
# shapefiles are in object mzkr
est<-Kernelheaping::dshapebivrProp(mzr2000_agg[,c(2:3,14:15)],burnin=b, samples=s,
adaptive=F,
                                shapefile=mzkr, gridsize=gs,
                                boundary=F,deleteShapes =
NULL,fastWeights=F,numChains = nc,numThreads = nt)
pdf("P:/MZKR-GEO/v2/schulweg_10-30min.pdf")
fields::image.plot(x=est$Mestimates$eval.points[[1]],
                   y=est$Mestimates$eval.points[[2]],
                   z=est$proportions,
                   col=RColorBrewer::brewer.pal(10,"RdYlGn"),
                   breaks=(0:10)/10,lab.breaks=(0:10)/10)
sp::plot(kreise,add=T)
title("Anteil Schulweg 10-30 min")
dev.off()
#-----
# way to school 30 minutes or more and less than 60 minutes
#-----
# Estimate using columns 2 and 3 as coordinates of centroids of MZ Kreisregionen
# and columns 16 as number of children with respective length of commute to scholl

```

```

# and columns 17 as number of all children
# shapefiles are in object mzkr
est<-Kernelheaping::dshapebivrProp(mzr2000_agg[,c(2:3,16:17)],burnin=b, samples=s,
adaptive=F,
                                shapefile=mzkr, gridsize=gs,
                                boundary=F,deleteShapes =
NULL,fastWeights=F,numChains = nc,numThreads = nt)
pdf("P:/MZKR-GEO/v2/schulweg_30-60min.pdf")
fields::image.plot(x=est$Mestimates$eval.points[[1]],
                   y=est$Mestimates$eval.points[[2]],
                   z=est$proportions,
                   col=RColorBrewer::brewer.pal(10,"RdYlGn"),
                   breaks=(0:10)/10,lab.breaks=(0:10)/10)
sp::plot(kreise,add=T)
title("Anteil Schulweg 30-60 min")
dev.off()
#-----
# way to school 60 minutes or more
#-----
# Estimate using columns 2 and 3 as coordinates of centroids of MZ Kreisregionen
# and columns 18 as number of children with respective length of commute to scholl
# and columns 19 as number of all children
# shapefiles are in object mzkr
est<-Kernelheaping::dshapebivrProp(mzr2000_agg[,c(2:3,18:19)],burnin=b, samples=s,
adaptive=F,
                                shapefile=mzkr, gridsize=gs,
                                boundary=F,deleteShapes =
NULL,fastWeights=F,numChains = nc,numThreads = nt)
pdf("P:/MZKR-GEO/v2/schulweg_gte60min.pdf")
fields::image.plot(x=est$Mestimates$eval.points[[1]],
                   y=est$Mestimates$eval.points[[2]],
                   z=est$proportions,
                   col=RColorBrewer::brewer.pal(10,"RdYlGn"),
                   breaks=(0:10)/10,lab.breaks=(0:10)/10)
sp::plot(kreise,add=T)
title("Anteil Schulweg gte 60 min")
dev.off()
#-----
# homosexual cohabiting couple
#-----
# Estimate using columns 2 and 3 as coordinates of centroids of MZ Kreisregionen
# and columns 22 as number of homosexual cohabiting couples
# and columns 23 as number of all cohabitong couples
# shapefiles are in object mzkr
est<-Kernelheaping::dshapebivrProp(mzr2000_agg[,c(2:3,22:23)],burnin=b, samples=s,
adaptive=F,
                                shapefile=mzkr, gridsize=gs,
                                boundary=F,deleteShapes =
NULL,fastWeights=F,numChains = nc,numThreads = nt)
pdf("P:/MZKR-GEO/v2/homo.pdf")
fields::image.plot(x=est$Mestimates$eval.points[[1]],
                   y=est$Mestimates$eval.points[[2]],
                   z=est$proportions,
                   col=RColorBrewer::brewer.pal(10,"RdYlGn")) # range adapted
sp::plot(kreise,add=T)
title("Anteil Gleichgeschlechtliche Kohabs")
dev.off()
#-----
# educational mobility downward movement
#-----
# Estimate using columns 2 and 3 as coordinates of centroids of MZ Kreisregionen
# and columns 24 as number of children with respective mobility
# and columns 25 as number of all children
# shapefiles are in object mzkr
est<-Kernelheaping::dshapebivrProp(mzr2000_agg[,c(2:3,24:25)],burnin=b, samples=s,
adaptive=F,
                                shapefile=mzkr, gridsize=gs,
                                boundary=F,deleteShapes =
NULL,fastWeights=F,numChains = nc,numThreads = nt)
pdf("P:/MZKR-GEO/v2/mobility1.pdf")
fields::image.plot(x=est$Mestimates$eval.points[[1]],
                   y=est$Mestimates$eval.points[[2]],

```

```

                                z=est$proportions,
                                col=RColorBrewer::brewer.pal(10,"RdYlGn"),
                                breaks=(0:10)/10,lab.breaks=(0:10)/10)
sp::plot(kreise,add=T)
title("Anteil Kind besucht nicht gym. Oberstufe, mind. 1 Elternteil hat Abi")
dev.off()
#-----
# educational mobility reproduction
#-----
# Estimate using columns 2 and 3 as coordinates of centroids of MZ Kreisregionen
# and columns 26 as number of children with respective mobility
# and columns 27 as number of all children
# shapefiles are in object mzkr
est<-Kernelheaping::dshapebivrProp(mzr2000_agg[,c(2:3,26:27)],burnin=b, samples=s,
adaptive=F,
                                shapefile=mzkr, gridsize=gs,
                                boundary=F,deleteShapes =
NULL,fastWeights=F,numChains = nc,numThreads = nt)
pdf("P:/MZKR-GEO/v2/mobility2.pdf")
fields::image.plot(x=est$Mestimates$eval.points[[1]],
                    y=est$Mestimates$eval.points[[2]],
                    z=est$proportions,
                    col=RColorBrewer::brewer.pal(10,"RdYlGn"),
                    breaks=(0:10)/10,lab.breaks=(0:10)/10)
sp::plot(kreise,add=T)
title("Anteil Kind besucht gym. Oberstufe und Eltern haben Abi oder beide nicht")
dev.off()
#-----
# educational upward movement
#-----
# Estimate using columns 2 and 3 as coordinates of centroids of MZ Kreisregionen
# and columns 28 as number of children with respective mobility
# and columns 29 as number of all children
# shapefiles are in object mzkr
est<-Kernelheaping::dshapebivrProp(mzr2000_agg[,c(2:3,28:29)],burnin=b, samples=s,
adaptive=F,
                                shapefile=mzkr, gridsize=gs,
                                boundary=F,deleteShapes =
NULL,fastWeights=F,numChains = nc,numThreads = nt)
pdf("P:/MZKR-GEO/v2/mobility3.pdf")
fields::image.plot(x=est$Mestimates$eval.points[[1]],
                    y=est$Mestimates$eval.points[[2]],
                    z=est$proportions,
                    col=RColorBrewer::brewer.pal(10,"RdYlGn"),
                    breaks=(0:10)/10,lab.breaks=(0:10)/10)
sp::plot(kreise,add=T)
title("Anteil Kind besucht gym. Oberstufe, kein Elternteil hat Abi")
dev.off()
#-----
# educational attainment no/Hauptschule
#-----
# Estimate using columns 2 and 3 as coordinates of centroids of MZ Kreisregionen
# and columns 34 as number of persons with respective educational attainment
# and columns 35 as number of all persons
# shapefiles are in object mzkr
est<-Kernelheaping::dshapebivrProp(mzr2000_agg[,c(2:3,34:35)],burnin=b, samples=s,
adaptive=F,
                                shapefile=mzkr, gridsize=gs,
                                boundary=F,deleteShapes =
NULL,fastWeights=F,numChains = nc,numThreads = nt)
pdf("P:/MZKR-GEO/v2/schulabschl_keinhaupt.pdf")
fields::image.plot(x=est$Mestimates$eval.points[[1]],
                    y=est$Mestimates$eval.points[[2]],
                    z=est$proportions,
                    col=RColorBrewer::brewer.pal(10,"RdYlGn"),
                    breaks=(0:10)/10,lab.breaks=(0:10)/10)
sp::plot(kreise,add=T)
title("Anteil höchster Schulabschluss kein Abschluss/Hauptschule")
dev.off()
#-----
# educational attainment Realschule / Polytechnische Oberschule
#-----

```

```

# Estimate using columns 2 and 3 as coordinates of centroids of MZ Kreisregionen
# and columns 36 as number of persons with respective educational attainment
# and columns 37 as number of all persons
# shapefiles are in object mzkr
est<-Kernelheaping::dshapebivrProp(mzr2000_agg[,c(2:3,36:37)],burnin=b, samples=s,
adaptive=F,
                                shapefile=mzkr, gridsize=gs,
                                boundary=F,deleteShapes =
NULL,fastWeights=F,numChains = nc,numThreads = nt)
pdf("P:/MZKR-GEO/v2/schulabschl_real_pos.pdf")
fields::image.plot(x=est$Mestimates$eval.points[[1]],
                   y=est$Mestimates$eval.points[[2]],
                   z=est$proportions,
                   col=RColorBrewer::brewer.pal(10,"RdYlGn"),
                   breaks=(0:10)/10,lab.breaks=(0:10)/10)
sp::plot(kreise,add=T)
title("Anteil höchster Schulabschluss Realschule/POS")
dev.off()
#-----
# educational attainment Abitur / Fachabitur
#-----
# Estimate using columns 2 and 3 as coordinates of centroids of MZ Kreisregionen
# and columns 38 as number of persons with respective educational attainment
# and columns 39 as number of all persons
# shapefiles are in object mzkr
est<-Kernelheaping::dshapebivrProp(mzr2000_agg[,c(2:3,38:39)],burnin=b, samples=s,
adaptive=F,
                                shapefile=mzkr, gridsize=gs,
                                boundary=F,deleteShapes =
NULL,fastWeights=F,numChains = nc,numThreads = nt)
pdf("P:/MZKR-GEO/v2/schulabschl_abi_fachabi.pdf")
fields::image.plot(x=est$Mestimates$eval.points[[1]],
                   y=est$Mestimates$eval.points[[2]],
                   z=est$proportions,
                   col=RColorBrewer::brewer.pal(10,"RdYlGn"),
                   breaks=(0:10)/10,lab.breaks=(0:10)/10)
sp::plot(kreise,add=T)
title("Anteil höchster Schulabschluss Abi/Fachabi")
dev.off()
#-----
# first nationality Turkish
#-----
# Estimate using columns 2 and 3 as coordinates of centroids of MZ Kreisregionen
# and columns 42 as number of persons with respective nationality
# and columns 43 as number of all persons
# shapefiles are in object mzkr
est<-Kernelheaping::dshapebivrProp(mzr2000_agg[,c(2:3,42:43)],burnin=b, samples=s,
adaptive=F,
                                shapefile=mzkr, gridsize=gs,
                                boundary=F,deleteShapes =
NULL,fastWeights=F,numChains = nc,numThreads = nt)
pdf("P:/MZKR-GEO/v2/nationality_tuerk.pdf")
fields::image.plot(x=est$Mestimates$eval.points[[1]],
                   y=est$Mestimates$eval.points[[2]],
                   z=est$proportions,
                   col=RColorBrewer::brewer.pal(10,"RdYlGn")) # Range adapted
sp::plot(kreise,add=T)
title("Anteil Türken")
dev.off()
#-----
# first nationality Italian
#-----
# Estimate using columns 2 and 3 as coordinates of centroids of MZ Kreisregionen
# and columns 44 as number of persons with respective nationality
# and columns 45 as number of all persons
# shapefiles are in object mzkr
est<-Kernelheaping::dshapebivrProp(mzr2000_agg[,c(2:3,44:45)],burnin=b, samples=s,
adaptive=F,
                                shapefile=mzkr, gridsize=gs,
                                boundary=F,deleteShapes =
NULL,fastWeights=F,numChains = nc,numThreads = nt)
pdf("P:/MZKR-GEO/v2/nationality_ital.pdf")

```

```

fields::image.plot(x=est$Mestimates$eval.points[[1]],
                  y=est$Mestimates$eval.points[[2]],
                  z=est$proportions,
                  col=RColorBrewer::brewer.pal(10,"RdYlGn")) # Range adapted
sp::plot(kreise,add=T)
title("Anteil Italiener")
dev.off()

#-----
# first nationality Former Yugoslavia
#-----
# Estimate using columns 2 and 3 as coordinates of centroids of MZ Kreisregionen
# and columns 46 as number of persons with respective nationality
# and columns 47 as number of all persons
# shapefiles are in object mzkr
est<-Kernelheaping::dshapebivrProp(mzr2000_agg[,c(2:3,46:47)],burnin=b, samples=s,
adaptive=F,
                                shapefile=mzkr, gridsize=gs,
                                boundary=F,deleteShapes =
NULL,fastWeights=F,numChains = nc,numThreads = nt)
pdf("P:/MZKR-GEO/v2/nationality_ex-jugo.pdf")
fields::image.plot(x=est$Mestimates$eval.points[[1]],
                  y=est$Mestimates$eval.points[[2]],
                  z=est$proportions,
                  col=RColorBrewer::brewer.pal(10,"RdYlGn")) # Range adapted
sp::plot(kreise,add=T)
title("Anteil Ex-Jugoslawen")
dev.off()

#-----
# first nationality Former GUS States
#-----
# Estimate using columns 2 and 3 as coordinates of centroids of MZ Kreisregionen
# and columns 48 as number of persons with respective nationality
# and columns 49 as number of all persons
# shapefiles are in object mzkr
est<-Kernelheaping::dshapebivrProp(mzr2000_agg[,c(2:3,48:49)],burnin=b, samples=s,
adaptive=F,
                                shapefile=mzkr, gridsize=gs,
                                boundary=F,deleteShapes =
NULL,fastWeights=F,numChains = nc,numThreads = nt)
pdf("P:/MZKR-GEO/v2/nationality_gus.pdf")
fields::image.plot(x=est$Mestimates$eval.points[[1]],
                  y=est$Mestimates$eval.points[[2]],
                  z=est$proportions,
                  col=RColorBrewer::brewer.pal(10,"RdYlGn"),
                  breaks=(0:10)/10,lab.breaks=(0:10)/10)
sp::plot(kreise,add=T)
title("Anteil Menschen aus GUS-Staaten")
dev.off()

#-----
# first nationality greece
#-----
# Estimate using columns 2 and 3 as coordinates of centroids of MZ Kreisregionen
# and columns 50 as number of persons with respective nationality
# and columns 51 as number of all persons
# shapefiles are in object mzkr
est<-Kernelheaping::dshapebivrProp(mzr2000_agg[,c(2:3,50:51)],burnin=b, samples=s,
adaptive=F,
                                shapefile=mzkr, gridsize=gs,
                                boundary=F,deleteShapes =
NULL,fastWeights=F,numChains = nc,numThreads = nt)
pdf("P:/MZKR-GEO/v2/nationality_greece.pdf")
fields::image.plot(x=est$Mestimates$eval.points[[1]],
                  y=est$Mestimates$eval.points[[2]],
                  z=est$proportions,
                  col=RColorBrewer::brewer.pal(10,"RdYlGn"),
                  breaks=(0:10)/10,lab.breaks=(0:10)/10)
sp::plot(kreise,add=T)
title("Anteil Griechen")
dev.off()

#-----
# first nationality poland

```

```
#-----
# Estimate using columns 2 and 3 as coordinates of centroids of MZ Kreisregionen
# and columns 52 as number of persons with respective nationality
# and columns 53 as number of all persons
# shapefiles are in object mzkr
est<-Kernelheaping::dshapebivrProp(mzr2000_agg[,c(2:3,52:53)],burnin=b, samples=s,
adaptive=F,
                                shapefile=mzkr, gridsize=gs,
                                boundary=F,deleteShapes =
NULL,fastWeights=F,numChains = nc,numThreads = nt)
pdf("P:/MZKR-GEO/v2/nationality_poland.pdf")
fields::image.plot(x=est$Mestimates$eval.points[[1]],
                  y=est$Mestimates$eval.points[[2]],
                  z=est$proportions,
                  col=RColorBrewer::brewer.pal(10,"RdYlGn"),
                  breaks=(0:10)/10,lab.breaks=(0:10)/10)
sp::plot(kreise,add=T)
title("Anteil Polen")
dev.off()
```

Code 3 Stata-Code zur Erstellung der 50 Bootstrap Samples des Männeranteils in Nordrhein-Westfalen

```

* Estimation of regional Means and Covariance Matrices of specific variables with MZ
* Regionalfile 2000
* Variables
* Geographical unit: MZKR (Mikrozensuskreisregion)
* Substantive Variables:
* Geschlecht

clear *
set more off
use ef711 ef32 using "R:\mzr\1506-
2010_Regionalfile_MZ_2000\STATA\mz00_regionalfile.dta", clear

* Nur Nordrhein-Westfalen
keep if inrange(ef711,5001,5999)

* generate proper ID
* gen SHN=string(ef711,"%05.0f")+ "00000"
gen double SHN=ef711*100000
order SHN
drop ef711

* Zufallsvariable
set seed 1
gen mann=ef32==1

local N=_N
* Create 50 Bootstrap-Samples
forvalues i=1/50 {
    preserve
    bsample `N'

    gen ding=mann!=.
    collapse (count) ding, by(SHN mann)
    drop if mann==.
    rename mann values
    rename ding mann_N
    bysort SHN: egen mann_NE=total(mann)

    reshape i SHN
    reshape j values
    reshape xij mann_N mann_NE
    reshape wide

    gen index=_n
    rename SHN SHN`i'
    drop mann_N0 mann_NE0
    rename mann_N1 mann_N1`i'
    rename mann_NE1 mann_NE1`i'
    compress
    sort SHN`i'
    save "P:\MZKR-GEO\bs`i'.dta", replace
    restore
}

* Merge 50 Bootstrap-Samples together
use "P:\MZKR-GEO\bs1.dta", clear
forvalues i=1/50 {
    merge 1:1 index using "P:\MZKR-GEO\bs`i'.dta", nogen noreport
}
order index
save "P:\MZKR-GEO\gender_bs.dta", replace

exit

```

Code 4 R-Code zur Schätzung und Visualisierung der Varianzen der räumlichen Verteilung des Männeranteils in Nordrhein-Westfalen

```

require(sp, maptools, rgdal, readstata13)

rm(list = ls())
setwd("P:/MZKR-GEO")
set.seed(1)
# Import Kreis-Shapes
kreise<-rgdal::readOGR("P:/MZKR-GEO/Verwaltungsgebiete1zu250000/vg250_2000-01-
01.utm32w.shape.ebenen/vg250_ebenen-historisch/de0001", "vg250krs")
# Keep only Northrhine-Westphalia
kreise<-kreise[substr(kreise$SHN,1,2)=="05",]
# Transform Kreis-Shapes to area-preserving projection, as Equal Area Grids based on
ETRS89-LAEA (EPSG 3035)
kreise<-sp::spTransform(kreise, sp::CRS("+init=epsg:3035"))

# Reproduce fusion of Kreise as was done in the Mikrozensus Regionalfile 2000
# -> Not necessary here, because all NRW-Kreise are MZ-Kreisregionen

# Sort Kreise in the order of the serial number SHN
kreise<-kreise[order(kreise$SHN),]
# Combine Kreise that have to be fused because of anonymity
mzkr<-maptools::unionSpatialPolygons(kreise, kreise$SHN)

# Merge MZ-Regionalfile 2000
# import stata file
genderbs<-readstata13::read.dta13("P:/MZKR-GEO/v2/gender_bs.dta",
                                con-
                                vert.factors=T, encoding="latin1", missing.type=T)

# put label points and shape ids from MZKR polygons in data frame
data<-
as.data.frame(cbind(sp::coordinates(mzkr), as.numeric(sp::getSpPPolygonsIDSlots(mzkr))
))
names(data)[names(data)=="V1"]<-"x"
names(data)[names(data)=="V2"]<-"y"
names(data)[names(data)=="V3"]<-"SHN"
# set parameters (5 burnin samples, 15 estimation samples, gridsize 100)
b<-5
s<-15
gs<-100
# loop over 50 bootstrap samples
for(i in 1:50) {
  # Merge MR-Kreis coordinates with
  # number of male persons and number of all persons across MZ-Kreisregionen
  # from bootstrap sample i
  eval(parse(text=paste("ct", as.character(i), "<-merge.data.frame(data, genderbs[,",
                        as.character((i-1)*3+2), ":", as.character((i-1)*3+4),
                        "], by.x=\"SHN\", by.y=\"SHN\", as.character(i),
                        "\", all.x=F, all.y=F)", sep="")))
  # Compute proportion for later visualisation, not necessary for estimation
  eval(parse(text=paste("ct", as.character(i), "$prop<-ct", as.character(i),
                        "$mann_N1_", as.character(i), "/ct", as.character(i),
                        "$mann_NE1_", as.character(i), sep="")))
  # Order with respect to MZ-kreis number
  eval(parse(text=paste("ct", as.character(i), "<-ct", as.character(i),
                        "[order(ct", as.character(i), "$SHN),]", sep="")))
  # estimate proportions with parameters (64 Chains, 8 Threads)
  eval(parse(text=paste("est", as.character(i), "<-Kernelheaping::dshapebivrProp(ct",
                        as.character(i), "[2:5], burnin=b, samples=s, adaptive=F,
shapefile=mzkr, gridsize=gs,
                        boundary=F, deleteShapes = NULL, fastWeights=F, numChains =
64, numThreads = 8)", sep="")))
}
# Estimate variance of proportions in MZ-Kreisregionen across bootstrap samples
ctbs<-ct1
names(ctbs)[6]<-"prop1"
for(i in 2:50) {

```

```

eval(parse(text=paste("part<-as.data.frame(ct",
                      as.character(i), "[,c(1,6)]", sep="")))
names(part)[1]<-"SHN"
eval(parse(text=paste("names(part)[2]<-\"prop\",
                      as.character(i), "\"\", sep="")))
ctbs<-merge.data.frame(ctbs,part,by.x="SHN",by.y="SHN",all.x=F,all.y=F)
}
ctbs$confint<-as.matrix(lapply(
  1:nrow(ctbs),function(i) stats::quantile(t(as.matrix(ctbs[i,6:8])),.95)
  -stats::quantile(t(as.matrix(ctbs[i,6:8])),.05)))
ctbs<-ctbs[order(ctbs$SHN),]
asd<-sp::addAttrToGeom(mzkr, ctbs, match.ID=F)
# plot of Variance estimate over MZ-Kreisregionen
pdf("P:/MZKR-GEO/gender_90percent_confint_choropleth.pdf")
sp::spplot(asd["confint"],
            col.regions=RColorBrewer::brewer.pal(10,"RdYlGn"),
            at=c((0:10)/10*.05),
            colorkey=list(at=c((0:10)/10*.25),
                          labels=as.character(c((0:10)/10*.25))))
dev.off()
# Estimate variance of proportions at each grid point across bootstrap sample
ding<-sapply(1:nrow(est1$proportions), function(i) {
  sapply(1:ncol(est1$proportions), function(j) {
    quan-
tile(sapply(1:50,function(k) eval(parse(text=paste("est",as.character(k),"$proportions
[j,i]",sep="")))),.95,na.rm=T)-
    quan-
tile(sapply(1:50,function(k) eval(parse(text=paste("est",as.character(k),"$proportions
[j,i]",sep="")))),.05,na.rm=T)
  })
})
# Plot Variance of proportions at each grid point across bootstrap sample
pdf("P:/MZKR-GEO/gender_90percent_confint_kdens.pdf")
fields::image.plot(x=est1$Mestimates$eval.points[[1]],
                  y=est1$Mestimates$eval.points[[2]],
                  z=ding,
                  col=RColorBrewer::brewer.pal(10,"RdYlGn"),

breaks=c((0:10)/10*0.25),lab.breaks=round(c((0:10)/10*0.25),digits=3))
sp::plot(kreise,add=T)
dev.off()

```
